Supporting document use through interactive visualisation of paragraph-level metadata

---

Mischa Weiss-Lijn

Ph.D. Computer Science

This thesis is submitted for the degree of Doctor of Philosophy of

University College London

December 02

Department of Computer Science

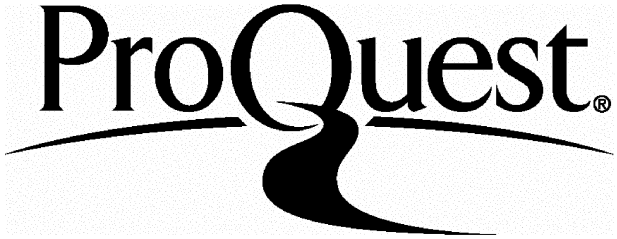
University College London

ProQuest Number: U642024

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest U642024

Published by ProQuest LLC(2015). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This work examines the thesis that the interactive visualisation of paragraph-level metadata can be used to increase the efficiency of goal-directed search in corporate documents. The design of a tool to support goal-directed search via the visualisation of paragraph-level metadata is explored, defined and implemented. A new evaluation method, named best-case analysis, is developed and applied to determine the potential utility of this tool.

A new visualisation, named GridVis, is described which displays a document's paragraph-level metadata. The design decisions are supported by task analysis, evaluation and a review of the information visualisation literature. A new technique based on interviews with users is developed to create metadata and metadata taxonomies that will service common information needs.

An interactive prototype of GridVis is subjected to an experimental performance evaluation. A qualitative analysis shows that performance could be improved if users made more use of the information provided by GridVis. The only quantitatively significant result was GridVis' lower task time which prompted suspicion of GridVis's utility. GridVis was re-implemented to realise the design ideas inspired by the qualitative analysis and user feedback.

A new evaluation method, named best-case analysis, is described. It makes use of GOMS and information foraging theory to build a model that computes a 'best-case' performance. This is then compared to the benchmark performance of twenty four domain experts using paper versions of the same documents. This evaluation method reveals that GridVis potentially offers either higher recall *or* faster task times than paper, but not both. It shows that the time saved by avoiding irrelevant paragraphs is

largely consumed by the extra navigation time required. If however the design were improved to make it 50% faster to use, it would give performance that is broadly superior to that of paper.

Contents

Abstract **2**

Contents **4**

Table of figures **13**

List of Publications **21**

Acknowledgments **23**

Chapter 1 Introduction **26**

1 Research questions and contributions *30*

2 Thesis structure *32*

Chapter 2 Background literature **36**

1 Introduction *36*

2 Metadata for use within documents *38*

2.1 Definition 38

2.2 Types of metadata 39

2.3 Producing metadata 40

2.3.1 Managing semantic heterogeneity 40

2.3.2 Structured documents 42

2.3.2.1 Mark-up languages for structured documents and metadata 42

2.3.2.2 The pros and cons of structured documents 44

2.4 Using metadata 45

2.4.1 Metadata as a finding aid 45

2.4.2 Dynamic document creation - microdocuments and metadata 47

2.5 Section summary 47

3 Information visualisation *49*

3.1 Layout techniques 49

3.1.1 Unstructured layouts 50

3.1.2 Structured layouts 51

3.1.2.1 Tree structures 52

3.1.2.2 Grid structures 54
3.2 Interaction techniques 58
3.2.1 Dynamic querying 58
3.2.2 Tightly coupled displays 58
3.2.3 Focus + context displays 60
3.3 Corpus visualisation 61
3.3.1 Visualisation of document metadata 61
3.3.2 Visualisation of category hierarchy 62
3.4 Single document visualisation 63
3.4.1 Visualisation of key word distribution in document surrogates 63
3.4.2 Visual document navigation and querying systems 67
3.4.3 Visualising document narratives 70
3.5 Section summary 72

4 Information seeking and information retrieval 74
4.1 Models of information seeking 74
4.2 The limitations of information retrieval techniques 76
4.3 Information seeking, document use and goal directed search in the workplace 78
4.3.1 Goal-directed search within-documents 79
4.3.2 The allure of paper 82
4.4 Section summary 84

5 Evaluation methods 85
5.1 Empirical techniques 85
5.1.1 Evaluation of within-document search tools 85
5.1.2 Evaluation of search result visualisations 88
5.1.2.1 Information retrieval metrics in visualisation evaluations 88
5.1.2.2 Alternative approaches to measuring performance 90
5.1.2.3 Difficulties of measuring search performance of visualisations 91
5.2 Theoretical evaluation techniques 91
5.2.1 General principles – rational analysis and cost structure 92
5.2.2 GOMS a mature engineering modelling tool 93
5.2.3 Other types of cognitive modelling 95
5.3 Section summary 96

6 Information foraging theory 97
6.1 Within and between patch foraging 100

6.2 The optimal diet selection algorithm 102
6.3 Constructing models of information foraging 105
6.4 Empirical support for information foraging theory 107
6.5 Section summary 110

7 Assumptions and limitations 111
7.1 A narrow focus on information retrieval in general and goal-directed search in particular 111
7.2 A focus on expert performance 113
7.3 A focus on traditional information retrieval metrics 115
7.4 A focus on quantitative comparative evaluation 116

8 Comparative review of related work 118
8.1 Visualisation 118
8.1.1 Navigation with visual document overviews 119
8.1.2 Navigation and search using visualizations of content distribution 120
8.1.3 Visualizations of metadata taxonomies 121
8.2 Hypertext and e-books 122
8.2.1 Outlining 123
8.2.2 Unstructured hypertext navigation techniques 124
8.3 Theoretical techniques for performance evaluation 126
8.4 Section summary 130

Chapter 3 The production and use of corporate documents and how metadata could fit in 133

1 Introduction 133

2 The writer study 135
2.1 Method 135
2.1.1 Methods used for the data analysis 137
2.2 Results and discussion 137
2.2.1 The document types produced at the Sainsbury's logistics planning division 138
2.2.2 Can goals be expressed as metadata? 139
2.2.3 Is the metadata replicating existing document organisation? 141
2.3 Summary 142

3 The reader study 143
3.1 Method 143
3.2 Results and discussion 145

3.2.1 The characteristics of long life, information rich documents 145
3.2.2 Reading practices 147
3.2.3 Characterising document reuse 148
3.2.4 Writer's understanding of the reader information needs 150
3.2.5 Can goals be expressed as metadata? 153
3.2.6 Is the metadata replicating existing document organisation? 154
3.3 Conclusion 154

Chapter 4 GridVis: a tool for supporting goal directed use through within-document metadata visualisation 157

1 Introduction 157

2 Creating Content-Descriptive Paragraph-Level Metadata for Business Documents 159
2.1 Selecting the documents 159
2.2 Metadata Design and Production 160
2.2.1 Tag Generation 160
2.2.2 Tag applicability 162
2.2.3 Tag Rationalization 164
2.3 Metadata Taxonomy Design and Production 165

3 Visualization Design 169
3.1 Design Methods 169
3.2 The Visualization Design 170
3.2.1 Overview 170
3.2.2 An Illustrative Scenario 172
3.2.3 Design Rationale and Advanced Features 173
3.2.3.1 Grid orientation 173
3.2.3.2 Navigation history 174
3.2.3.3 Paragraph context 174
3.2.3.4 Making the grid scannable 175
3.2.3.5 Supporting the use of tags 176

4 Alternative design features considered 181

5 System Implementation 183
5.1 Encoding the metadata in XML 183
5.2 The visualisation implementation 185

6 Chapter summary 186

Chapter 5 Initial evaluation 187

1 User trial – Exploring the capabilities of GridVis 187

1.1 Method 188

1.1.1 Data analysis 192

1.2 Results 193

1.3 Discussion 194

2 Task analysis 197

2.1 Objects and the object taxonomy 198

2.2 Actions, the goal structure and the procedural substructure 200

2.3 Key goals and objects: a simplified task model 204

2.4 Discussion 206

3 Experimental evaluation 207

3.1 Method 207

3.1.1 The sample 208

3.1.2 The GridVis prototype and technical set-up 208

3.1.3 The Documents 211

3.1.4 Relevance ratings 212

3.1.5 Experimental design 214

3.1.6 Data capture 216

3.2 Quantitative results 216

3.2.1 Analysis 216

3.2.2 Retrieval performance results 218

3.3 Qualitative results 219

3.3.1 Investigating strategy use 219

3.3.1.1 Analysis procedure 219

3.3.1.2 Tag cell selection strategies 222

3.3.1.3 Paragraph cell selection 224

3.3.2 Investigating the use of co-occurrence information in cell selection ... 228

3.3.2.1 Analysis procedure 228

3.3.2.2 Results – an exploration of the effect of paragraph cell selection strategy 230

3.4 Discussion 232

4 The problems of empirical evaluation in a corporate environment 236

5 Conclusions - Implications for the further evaluation of GridVis 238

Chapter 6 A best-case analysis evaluation of the potential of within-document metadata for goal-directed search 240

1 Introduction 240

1.1 Best-case analyses 241

1.1.1 What is a best-case analysis? 241

1.1.2 The pros and cons of best-case analyses 243

1.1.3 Approaches to constructing best-case analyses 245

1.2 Applying the best case evaluation approach to GridVis 246

2 Constructing the best case model 249

2.1 Strategy level model definition 250

2.1.1 Paragraph cell scent evaluation strategy 257

2.1.2 Tag scent evaluation strategy 259

2.1.3 Assumptions used in strategy level model 260

2.1.4 Summary of the strategy level model 262

2.2 Performance level model definition 263

2.2.1 Currency Assumptions 266

2.2.2 Scent models 268

2.2.2.1 Tag information scent model 268

2.2.2.2 Paragraph cell information scent model 270

2.2.3 Cost models 275

2.2.3.1 Models of actual cost 275

2.2.3.1.1 Tag choice cost model 277

2.2.3.1.1.1 Sanity check experiment 281

2.2.3.1.2 Paragraph cell choice cost model 282

2.2.3.1.2.1 Sanity check experiments 287

2.2.3.1.3 Paragraph handling model 292

2.2.3.2 Models of expected cost 297

2.2.3.2.1 Estimating GridVis navigation time 299

2.2.3.2.2 Estimating paragraph handling time 308

2.2.3.2.3 Estimating the mean patch choice time 311

2.3 Section summary – an overview of the best-case model 311

3 Paper performance experiment 313

3.1 Sample 313

3.2 Materials 314

3.2.1 The documents 314

3.2.1.1 Choice of document type 315
3.2.1.2 Varying document length 316
3.2.1.3 Varying information need difficulty 317
3.2.2 Developing representative queries 318
3.2.3 Relevance judgements 319
3.3 Method 319
3.3.1 Data transcription 322
4 Performance Metrics 324
5 Chapter summary 326
Chapter 7 Results from the best-case analysis 327
1 The benchmark – performance and strategies with paper 327
1.1 Reading strategies 327
1.2 Task performance 333
1.3 The effect of document length 335
1.4 The effect of question difficulty 338
2 How does the best-case GridVis performance fare against the benchmark? 340
2.1 Recall performance 341
2.2 Precision performance 344
2.3 Task time performance 345
2.4 Does the extra cost of navigation with GridVis outweigh the cost savings of the paragraphs avoided? 346
2.5 Does GridVis increase the quality and relevance of material found? 348
2.6 The trade-off – balancing task time and retrieval effectiveness 349
2.7 The effect of document length 350
2.8 The importance of exploitation cost prediction 351
3 General discussion 353
3.1 Design implications 355
3.2 Critical assessment 355
4 Chapter summary 358
Chapter 8 Conclusions and further work 360
1 Conclusions and contributions 360

1.1 A new approach for supporting goal-directed search within corporate documents ... 360
1.2 A new evaluation technique ... 363
1.3 Evaluation of the functional viability of interactive visualisations of paragraph-level metadata ... 364

2 Further work ... 366
2.1 Evaluation – alternative scenarios of usage ... 366
2.2 Paragraph-level metadata and its visualisation ... 367
2.3 Information foraging models ... 368

References 371

Appendix 1 – Results and discussion for the overview and sorting task in the user study ... 384

1 Discussion ... 385

Appendix 2 – Expert judge ratings ... 387

Appendix 3 - Configurations of tag selection strategies ... 390

Appendix 4 – GOMS models for tag and paragraph cell selection ... 393

2 GOMS model for tag selection in the GridVis Interface ... 393
2.1 High level goals ... 393
2.2 Methods ... 393

3 GOMS model for Paragraph cell selection times ... 396
3.1 Left to right strategy ... 396
3.2 Applicability only ... 396
3.3 Co-occurrence only ... 397
3.4 Co-occurrence and restricted applicability ... 399

4 Operator algorithm for get-from-task <cell-applicability> ... 401

Appendix 5 - Paragraph time coding procedure ... 403

1 Codes assigned to paragraphs ... 405

Appendix 6 – Graphs showing full data for the effect of document length in the paper performance experiment ... 407

Appendix 7 Further details on the documents used in the best-case analysis ... 413

1 *Setting and measuring document length* *413*

2 *Other considerations in the document construction* *414*

Table of figures

Figure 1 Thesis Map. The location of each major and minor contribution 33

Figure 2 SemNet (Fairchild, Poltrock & Furnas 1988). The image on the left shows a network with random placement of nodes while the other shows it once the graph layout has been structured. 50

Figure 3 Cone Tree (Robertson, Mackinlay & Card 1991) 53

Figure 4 Overview of teaching resources (Marchionini & Shneiderman 1988) 56

Figure 5 The Matrix Browser (Botsch & Kunz 1996) 57

Figure 6 PDQ Tree-browser (Kumar, Plaisant & Shneiderman 1997) 59

Figure 9 The PopOut Prism (Suh et al. 2002) 66

Figure 10 The Document Lens (Mackinlay & Robertson 1993) 68

Figure 12 Three document visualisations (Hornbæk & Frøkjær 2001) 69

Figure 13 A general model of information seeking (Hearst 1999) 75

Figure 14 The prevalence of goal-directed, and non goal-directed types of reading, adapted from Adler et al (Adler et al. 1998) 80

Figure 15 Searching in a digital library; characterised as a repeating cascade of patches and items 101

Figure 16 An example of the relationship of between item profitability and the profitability of diets including items 1,2, ... k 105

Figure 17 This graph shows that the productions for selecting clusters that were more highly ranked by ACT-IF, were more often selected by users. 108

Figure 18 On the y-axis, this graph shows the probability of the next cluster not being selected, subtracted from probability of it being selected. On the x-axis, it shows the profitability of the next item, minus the profitability of the current diet. More simply, whether a cluster will be selected is shown on the y-axis, and whether the profitability for the current item is larger than that for those already selected is shown on the x-axis..... 109

Figure 19 The proportion of goals with each level of diffuseness and evaluativeness rating. The portions of the graph that are hashed out indicate those goals which are not metadata expressible. 141

Figure 20 Recall and precision, for the goals predictions made by writers of four documents (one reader/participant for each document) 151

Figure 21 Recall and precision for predictions of the portions of text needed by reader 152

Figure 22 The proportion of goals with each level of diffuseness and evaluativeness rating. The hashed portions of the graph indicate those goals which are not metadata expressible. 154

Figure 23 The components of the GridVis visualisation 171

Figure 24 The paragraph cells relating to the tag selected are given a different highlight colour, allowing the user to locate them effortlessly. 176

Figure 25 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold),the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines) 177

Figure 27 The structure of the XML documents visualised by GridVis. (Dotted lines indicate optional elements)..... 184

Figure 28 Screen shot of prototype GridVis mock-up used..... 190

Figure 29 Evaluation interface for the goal-directed search task..... 191

Figure 30 The object taxonomy..... 200

Figure 32 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold),the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines) 205

Figure 33 A screenshot of the version of GridVis used in the experimental evaluation 209

Figure 34 GridVis as used in the experiment..... 210

Figure 35 Screenshot of the setup for the 'without GridVis' condition. 211

Figure 36 The mean scores for each tag selection strategy across all participants, show as percentage of total score..... 223

Figure 37 The mean scores for each paragraph cell selection strategy across all participants show as percentage of total score..... 224

Figure 38 Scores for each paragraph cell selection strategy across three participants with a similar configuration of strategies, shown as percentage of total score..... 225

Figure 39 Scores for each paragraph cell selection strategy across six participants with a similar configuration of strategies, shown as percentage of total score..... 226

Figure 40 Interactive recall and traditional precision for the top ranked 3rdof paragraph cells using each of 3 different paragraph selection strategies. 231

Figure 42 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold), the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines) 251

Figure 43 An overview of the best-case model 252

Figure 44 The use of GridVis seen as a cascade of patches and items 255

Figure 45 Illustration of the three different types of information available for judging the information scent of paragraph cells 258

Figure 46 Overview of GridVis best-case model 262

Figure 47 Overview of the tag choice portion of the performance level model 265

Figure 48 Overview of the paragraph cell choice portion of the performance-level best-case model 266

Figure 49 A scatter plot showing the correspondence between information value predictions based on applicability alone and the actual information values from the expert judge relevance ratings 272

Figure 50 A scatter plot showing the correspondence between information value predictions based on co-occurrence alone and the actual information values from the expert judge relevance ratings 273

Figure 51 A scatter plot showing the correspondance between information value predictions based on applicability and co-occurrence and the actual information values from the expert judge relevance ratings 275

Figure 52 An illustration of how the actual cost models, , are broken down, and what data, , is used to set their parameters 276

Figure 53 The basic structure of the tag choice NGOMSL model 277

Figure 54 Chart illustrating the discounting of tag choice cost each time a section is revisited 279

Figure 55 Human and GOMS times for tags choice. The trials are shown in sequential order; thus, repeated section tags indicate section revisits 282

Figure 56 Illustration of the cost model for the LtoR strategy 283

Figure 57 Ilustration of the cost model for the AppOnly strategy 284

Figure 58 Ilustration of the cost model for the CoocOnly strategy 285

Figure 59 Ilustration of the cost model for the Cooc&RestrictedApp strategy 285

Figure 60 Human and GOMS cell choice times using the LtoR strategy 289

Figure 61 Human and GOMS cell choice times using the AppOnly strategy 290

Figure 62 Human and GOMS cell choice times using the CoocOnly strategy 291

Figure 63 Human and GOMS cell choice times using the Cooc&RestrictedApp strategy 292

Figure 64 Scatter plot of paragraph length against predicted and actual times (from the paper performance experiment) for paragraphs that are ultimately accepted 294

Figure 65 Scatter plot of paragraph relevance rating (i.e. information value) against predicted and actual times (from the paper performance experiment) for paragraphs that are ultimately accepted 295

Figure 66 Scatter plot of paragraph length against actual and predicted times for reading paragraphs that are utlimately rejected 296

Figure 67 Scatter plot of paragraph relevance (i.e. information value) against actual and predicted times for reading paragraphs that are ultimately rejected 296

Figure 68 Illustration of the different processes involved in section and tag exploitation cost 298

Figure 69 An illustration of the components of expected cost, ▭, and what data, ▱, is used to set their parameters 299

Figure 70 An illustration of how the tag selection cost is broken down into various components, ▭, and what data, ▱, is used to set their values and accompanying coefficients and constants ○ 300

Figure 71 An illustration of how the estimation section selection cost is broken down into various components, ▭, and what data, ▱, is used to set their values and accompanying coefficients and constants ○ 304

Figure 72 The percentage difference between the predicted and 'actual mean number of unvisited paragraph cells' (shown as a percentage of the actual number of unvisited cells). The cross hair drawn with a dotted line, illustrates that the error tends to 0 when 'actual mean number of cells on a tag' tends to, 3, its mean. 307

Figure 73 A query with two constraints on relevance, indicated by the grey panels. .. 317

Figure 74 The collated data on reading strategies and the transitions between them, from interview and observational data. Transitions that were only evidenced by one subject have been excluded to clarify the overall picture. By starting at the top and tracing a path downwards, one can appreciate the use of multiple layers of reading strategy, which increase in cost as the expected relevance of material increases. 331

Figure 75 The mean time taken to complete a question for different document lengths. 336

Figure 76 The mean result and traditional precision, and 'proportion of reading time spent reading relevant' across the different document lengths..... 337

Figure 77 The mean recall across the different document lengths. 338

Figure 78 Overall Recall scores 342

Figure 79 Core and Peripheral recall scores..... 343

Figure 80 Precision performance scores..... 344

Figure 81 Task time performance..... 345

Figure 82 The dependence of task time on the cost estimates for GridVis usage. The results for paper are not part of the sensitivity analysis, but are included to enable comparison..... 346

Figure 83 Scores for the proportion of time spent reading and navigating..... 347

Figure 84 Results from the sensitivity analysis of the proportion of times spent reading and navigating..... 348

Figure 85 The rate of gain or profitability (information units per second)..... 349

Figure 86 Results from the sensitivity analysis for the rate of information gain..... 350

Figure 87 An illustration of when paper or GridVis is advantageous..... 354

Figure 88 Scores for each tag selection strategy across four participants with a similar configuration of strategies, shown as percentage of total score..... 390

Figure 89 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score..... 391

Figure 90 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score..... 391

Figure 91 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score..... 392

Figure 92 The individual participant times and the mean scores, for core recall against document length..... 407

Figure 93 The individual participant times and the mean scores, for peripheral recall against document length 408

Figure 94 The individual participant times and the mean scores, for overall recall against document length 409

Figure 95 The individual participant times and the mean scores, for task time against document length 409

Figure 96 The individual participant times and the mean scores, for interactive precision against document length 410

Figure 97 The individual participant times and the mean scores, for traditional precision against document length 411

Figure 98 The individual participant times and the mean scores, for proportion of reading time spent with relevant against document length 412

Figure 99 The individual participant times and the mean scores, for task time against document length 412

List of Publications

An early version of Chapter 4 has been published in:

> Mischa Weiss-Lijn, Janet T. McDonnell, Leslie James (2002) Interactive Visualization of paragraph-level metadata to support corporate document use. Visualising the Semantic Web, Vladimir Geroimenko and Chaomei Chen (Eds.) Springer Verlag, Nov 2002

Portions of chapter 4 have also appeared in:

- Mischa Weiss-Lijn (2002) Visualisation de métadonnées pour facilité l'utilisation des documents. IHM'2002, Poitiers

The experimental evaluation reported in section 3 of chapter 5 has been published in:

- Mischa Weiss-Lijn, Janet T. McDonnell, Leslie James (2002) An empirical evaluation of the interactive visualization of metadata to support document use. Lecture Notes in Computer Science – Visual Interfaces to digital libraries (in press)

This evaluation, along with portions of chapter 4, was also summarised in:

- Mischa Weiss-Lijn, Janet T. McDonnell, Leslie James (2001) Supporting document use through interactive visualization of metadata. Visual Interfaces to Digital Libraries - Its Past, Present, and Future, ACM+IEEE Joint Conference on Digital Libraries, June 28th 2001

These parts of the thesis are also covered, along with a discussion, from section 4 of chapter 5, on the difficulties of doing empirical evaluations in the workplace, in:

- Mischa Weiss-Lijn, Janet T. McDonnell, Leslie James (2001) Visualising Document Content with Metadata to Facilitate Goal-directed Search. International Symposium of Visualisation of the Semantic Web, IV2001 25 - 26 - 27 July 2001

An early formulation of the best-case analysis approach, and its application to GridVis, is described in chapter 6 has been published in:

- Mischa Weiss-Lijn, Janet T. McDonnell, Leslie James (2001) A 'best-case' evaluation of novel information retrieval applications. 8th European Conference on Information Technology Evaluation 17-18 September 2001

Acknowledgments

A great many people have at some stage been involved in work that has found its way into this thesis.

First and most formally, this research was sponsored by J Sainsbury plc and undertaken within the Postgraduate Training Partnership established between Sira Ltd and University College London. Postgraduate training partnerships are a joint initiative of the DTI and the EPSRC.

The work in this thesis was supported by and carried out in three institutions, Sainsbury's, Sira, and UCL.

At Sainsbury's, the energy and enthusiasm of the now disbanded knowledge group, inspired and fuelled the first year and a half of research. Marc Venables, Zahir Bashir, Ian Hawkins, Stuart Robinson all took an active interest in my work and helped form my early ideas. Duncan Wilson, my guardian and mentor at Sainsbury's, was instrumental in enabling me to involve the people and problems of Sainsbury's in the research. His diligence in protecting me from the frequent waves of corporate reorganisation was such that I am the only member of the knowledge group to survive them. When the knowledge group was disbanded, Duncan left me in the capable hands of Leslie James. Leslie, and his Strategy Research Team (referred to later as the 'frontier users'), provided a me with a huge amount of support. Leslie, Clair, Dave, and Rob, were all repeatedly willing to take time out to act as volunteers for the various experiments, user tests, and metadata production tasks which needed to be done by domain experts. Leslie in particular, through his intelligence, resourcefulness and keen

sense of good design, contributed enormously. Finally, thanks are due to the many people at Sainsbury's head office who kindly agreed to be subjects in my experiments.

Sira provided me with a place to work, free lunches, vocational training, and most importantly of all, the wise counsel of Dr John Gilby. Despite a very busy schedule, John was always willing to give me the time I needed, fearlessly diving into the details of any mathematical and algorithmic tangles I'd got myself into. His advice was always practical but encouraging, keeping me on the narrow winding path to completion.

At UCL I started life in the Virtual Environments research group under the tutelage of Mel Slater. He guided me through the first year and a half with insightful critiques and a wealth of ideas. The Virtual Environments group provided me with a congenial and welcoming base. Manuel Oliveira was the exemplary PhD student with whom I shared a desk. He always took an active and constructive interest in my work, giving the harsh but true critiques that I needed to hear.

When my work moved out of Mel Slater's domain, Janet McDonnell bravely took me on as her PhD student. She has always been there, to help clarify problems thrown up by the research, and to untangle the narratives in my writings. Janet guided me through the subtlies of doing research in a corporate context and helped me kick the 'modelling' habit enabling me to accept that the best-case model was finally finished.

Once under Janet's wing, I moved to the multimedia lab. Here I was able to discuss my work and the information seeking literature with Simon Attfield who was one of the very few people around whose research interests overlapped with mine. John McCarthy, my former employer and desk mate, provided invaluable advice on statistics. I'd also like to thank the rest of the lab for helping with momentous task of proof reading my thesis. John Dowell and Angela Sasse, HCI lecturers at UCL, kindly took time out on

several occasions, to give me invaluable advice on aspects of HCI and research methodology.

I must of course thank Anne Poméry, my darling girlfriend, who never managed to understand why I was researching in this area, but always gave me the love and affection I needed. Finally there is my (nuclear) family who were almost able to explain what my thesis was about, after just four years. They housed me and fed me once my grant money ran out, and even proof read some of my thesis; bless their cotton socks.

Chapter 1 Introduction

Large corporations have for many years attempted to improve the way they use, manage and distribute information and knowledge. Among these attempts, a number of technologies have been introduced to support the use of documents. Documents are a major repository of corporate knowledge; their use is woven into the fabric of office work. Thus, improvements in the tools for document use can contribute directly to increased productivity.

In general, currently used technologies, offer incomplete support for document use. For example, many companies have recently deployed intranet systems to more effectively store and distribute documents (IDC 2001). These technologies offer incomplete support for document use because they focus on document delivery. This is problematic because a substantial portion of document use consists of goal-directed searches where only a portion of the document is required. Thus although the digital support stops once a desired document is found, the search does not. This suggests that there is an opportunity to improve matters by extending the digital support for search to within-document search.

Digital support for within-document search could have a substantial impact despite the widespread use of paper for reading documents. As early dreams of the paperless office have withered, it has been argued that paper persists because it offers important affordances[1] that are difficult to reproduce in electronic form (Sellen & Harper 1997). Nevertheless, on-screen reading of documents is common; in two studies of workplace

reading (Adler et al. 1998; Sellen & Harper 1997) document related tasks accounted for between 97% and 82% of total work time. In one of the studies just under 50% of this time was spent with a combination of paper and electronic documents (Sellen & Harper 1997). Both studies found that about 13% to 14% of the time was spent with digital documents alone. Thus, tools which improve the efficiency of on-screen document use could be of value to corporations.

Documents are used in a wide variety of ways; the research presented in this thesis focuses on goal-directed search which is an important and prevalent class of document use. Goal-directed search can be defined as "sampling information in the text which satisfies the goal of the search" (Adler et al. 1998, p. 244). Documents are used for goal-directed search, cross-referencing, skimming, prolonged linear reading, annotation (Adler et al. 1998), and even used as tangible symbols to convey social messages (Sellen & Harper 1997). They are integral to many different types of work such as writing, analysis, discussion and collaboration. This research focuses on attempting to support goal-directed search because it is part of a variety of important work activities such as writing and analysis, and is itself a prevalent workplace activity.

This research examines one of a number of possible avenues (Sellen & Harper 1997) for improving the digital support of document use. One approach, currently pursued by the e-book community, is to develop new hardware that reproduces some of paper's affordances such as legibility, portability and annotatability (Schilit, Price & Golovchinsky 1998). Alternatively, software tools can be developed to better support document use with standard hardware. The use of string searching, e.g. a 'find in page'

---

[1] The term affordance refers to the perceived and actual properties of an artefact which inform the user as to its function and mode of use. For example, a chair affords ("is for") support, and, therefore, affords sitting.

function, is a typical widely used example. Similarly outlining is sometimes used to show document structure and content via a tree view of headings. More recently, researchers have attempted to apply techniques from the field of information visualisation to support document use. For example, visualisation has been used in combination with string searching to visualise the distribution of the keywords being sought. Another approach has been to generate visual thumbnails which convey large-scale features such as section and paragraph boundaries and illustrations. Another, new, approach is explored and evaluated in this thesis; the interactive visualisation of paragraph-level metadata.

The interactive visualisation of paragraph-level metadata is an attractive avenue to explore for a number of reasons. Firstly, paragraph-level metadata can give accurate descriptions of content which are tailored to the interests of its users. It can do so using the local language of the users, thus reducing problems introduced by variations in vocabularies used (Furnas et al. 1987). These descriptions can be aggregated at multiple levels of both structure and content. A user can examine the content at the level of any aggregate of paragraphs present in the document such as sub-sections or section. Thus, they can work at a level of granularity appropriate to their needs. If the metadata sits in a taxonomy, the user can also look for content related to categories which collect together numerous related pieces of metadata. Finally, in combination with the use of information visualisation techniques this paragraph-level metadata can be used to show the presence, location and extent of content in a document.

Although the development of new technologies is in itself valuable, their evaluation is perhaps more so. Evaluation is crucial in both practical and academic terms. It is needed to determine the potential for actual deployment of a new technology. If a tool is to be used in a business environment, there must be reason to believe that it will confer

significant advantages. This is especially true for the use of paragraph-level metadata, where there will be a significant cost associated with production of the metadata. Evaluation is also needed in order to make a complete and useful contribution to the field of human-computer interaction (HCI). Without evaluation, the utility of any avenues explored is uncertain. It would be unclear whether these avenues can be can safely ignored by future research, or profitability pursued.

In HCI research, what is of most value is the potential of the class of designs being explored through the development of a particular system. Minor design problems, the shortcomings of a prototype implementation or the inexperience of users can be overcome by further design, development work and training if there is sufficient potential to warrant it. It is thus valuable to be able to directly assess the potential of a design. Since interactive software rarely meets completely novel goals, its potential can often usefully be assessed by comparing its potential performance to that of a benchmark, i.e. an existing means to achieve the same goals. In this research, a new evaluation methodology, best-case analysis, is developed and demonstrated, which allows the potential of a class of designs to be assessed relative to a benchmark tool.

1 Research questions and contributions

The discussion above leads to three research questions:

1. How can the interactive visualisation of paragraph-level metadata be used to support goal-directed search, within-documents, in a corporate environment?
2. Can an interactive visualisation of paragraph-level metadata be used to deliver worthwhile support of goal-directed search, within-documents, in a corporate environment?
3. How can the potential of a class of designs be evaluated, rather then the performance of a particular implementation?

These in turn result in three major contributions and a number of minor contributions. The major contributions are:

An approach to visualising within-document metadata which facilitates goal-directed search within-documents

A new evaluation method which allows the ultimate potential of a class of designs to be assessed

An evaluation of the utility of the interactive visualisation of paragraph-level metadata in the support of goal-directed search within-documents

The minor contributions are:

A characterisation of the documents suitable for enrichment with paragraph-level metadata

A new method for soliciting metadata and metadata taxonomies that are focused on the information needed to discriminate relevant material

The application of a best-case analysis to an information seeking tool

A characterisation of the high-level strategies used in goal-directed reading

The uncovering of theoretical grounds, in information foraging theory, that suggest the importance, for information seeking tools, of cues which communicate information on access costs.

2 Thesis structure

To convey the structure of this thesis, each chapter is described very briefly at first and summarised by the thesis map in Figure 1, the contents of each chapter are then described in more detail. Chapter 2 reviews the literature in the fields related to this work. It describes relevant background material and lays out the evidence to warrant the choice of techniques used later in the thesis for design and evaluation work. Chapter 3 reports two exploratory studies of document use in a corporate environment which sought to assess the feasibility of using paragraph level metadata. Chapter 4 describes the design of GridVis, an interactive visualisation of paragraph-level metadata to support goal-directed search. Chapter 5 reports user trials, task analysis, and experimental evaluation that were used to assess and further refine the design of GridVis. Chapter 6 describes the best-case analysis used to ascertain the potential performance of the interactive visualisations of paragraph-level metadata as embodied by GridVis. Chapter 7 reports the results of this evaluation, and Chapter 8 ties together the conclusions and contributions of this research and suggests avenues for further work. Figure 1 illustrates the way each of these chapters fit into the thesis as a whole, and locations of the various contributions enumerated in section 1. The content of each chapter will now be described in more detail.

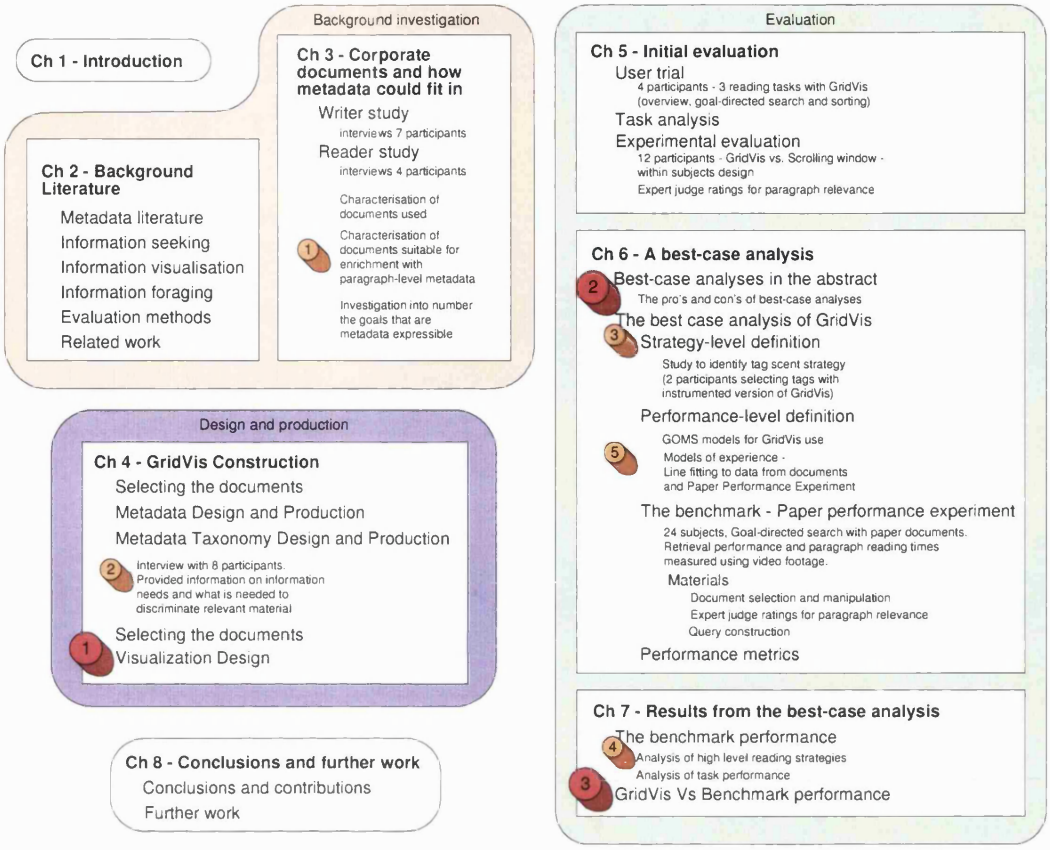


Figure 1 Thesis Map. The location of each major 1 and minor 1 contribution

Chapter 2 reviews literature from the various areas related to this research. Section 2 on metadata describes document metadata from different traditions, looks at some of the issues surrounding its production, and reviews the ways it has been used to support search within documents. A section on information visualisation, section 3, reviews its various techniques and fields evidence to warrant the design choices in the development of GridVis. The following section, section 4, draws on theories of information seeking to argue for the potential utility of metadata taxonomies, and uses evidence for the importance of goal-directed search in the workplace to argue that it is worth supporting and notes the continuous prevalence of paper document use. The next section, 5, looks at evaluation methods. It examines the empirical methods and metrics used to evaluate the support of goal-directed search within documents. A range of theoretical evaluation

methods are described and the particular choice of techniques used in this research is justified. One of the more recently developed of these techniques, the use of information foraging theory, is the subject of the following section, 6, where its foundations, use and validity are discussed. The final section, section 8, covers visualisation work related to the work in this thesis, it is critically assessed and the inadequacies of evaluation techniques which might substitute for the best-case method developed in this research, are exposed.

Chapter 3 reports two exploratory studies of document use in the head office of a large UK retailer which attempted to examine the feasibility of using paragraph-level metadata to support goal-directed search within corporate documents. A subset of the documents used are identified as being suitable for enrichment with metadata. Their characteristics are described. An analysis of the goals workers bring to documents they read examines the degree to which these goals can be represented with metadata made up of keywords or noun phrases.

Chapter 4 describes the design and construction of GridVis, an interactive visualisation of paragraph-level metadata. The techniques used to construct the high quality metadata needed are described in section 2. These include techniques adapted from previous research and a novel technique developed in this research, to elicit the information used to discriminate relevant material. The design of GridVis is described in section 3; the contribution to the design, of evidence from numerous rounds of evaluation, is discussed.

Chapter 5 reports a user trial, task analysis and experimental evaluation of GridVis. In section 1, the description of a user trial of an early static mock-up of GridVis supports discussion of a number of design problems and highlights methodological concerns. A task analysis of the use of paragraph-level metadata for goal-directed search is then

reported for use in a review of the GridVis design (see section 2). The final section of this chapter, section 3, describes the evaluation of a full implementation of an improved design, using a controlled experimental comparison with a simple scrolling window. Results from this experiment are used to argue for the need for an evaluation methodology which can assess the ultimate potential of a class of designs.

Chapter 6 describes the best-case analysis method and its application to the interactive visualisation of paragraph-level metadata. The best-case analysis method is described in section 1.1. Its general advantages and limitation are discussed. In section 2 the best-case model of GridVis use which incorporates information foraging theory is described. First, in section 2.1, the model is described at a high, or strategy level. Then, in section 2.2, it is described at a computational, or performance level. Section 3 describes an experiment to measure human goal-directed search performance with paper documents.

Chapter 7 reports the results of the best-case analysis. The results from the experiment on performance with paper are reported first in section 1. The navigational strategies used with paper are described and used to argue the validity of comparing the performance on paper with that computed by the best-case model of GridVis. The best-case performance of GridVis is compared with that of paper in section 2. The relative advantages of various affordances offered by GridVis are reported and an assessment is made of the general viability of interactive visualisations of paragraph-level metadata.

Chapter 8 summaries the contributions and conclusions offered by this research and describes various avenues for further work.

Chapter 2 Background literature

1 Introduction

In this chapter literature from various related fields is reviewed to portray the context of the research, lay out the evidence behind the direction of this work, explain the theoretical and methodological tools used, gather together evidence used to warrant the research and design decisions made and critically evaluate related work.

A section on metadata, section 2, lays out the different traditions in which metadata has been used, and describes some systems which have used it at a sub-document level. It covers the various technical standards and tools with which paragraph-level metadata can be produced and discusses problems of metadata production and their associated solutions. The following section, section 3, on information visualisation, reviews the various techniques used in the design work of this thesis, and fields evidence to warrant the design choices made. Next, section 4, uses the theoretical work on information seeking to argue for the potential utility of metadata. It goes on to review standard information retrieval techniques, and their suitability for searching within documents. Finally, it reviews studies of workplace document use to argue the importance of goal directed search, and the suitability of paper as a benchmark against which technologies for within-document search can be compared. The following section, section 5, on evaluation methods reviews the range of empirical and theoretical techniques that have been, or could be, applied to document search tools. Lessons are drawn for the empirical evaluation carried out in this research and the choice of theoretical techniques is justified against a background of alternatives. In section 6, one of the more novel theoretical tools used, information foraging theory, is described in some depth, in order

to prepare the reader for its use in the evaluation work described in Chapter 6. Finally section 8, surveys related work, looking at previous attempts to support within document search and existing evaluation techniques contrasting them with the tool and evaluation technique presented in this thesis.

2 Metadata for use within documents

This section looks at document metadata, since the research presented in this thesis explores the potential of paragraph-level metadata that describes the semantic content of corporate documents.

The section starts by distinguishing between metadata arising from the library cataloguing tradition, and that coming from the artificial intelligence and knowledge based systems community. Turning to the production of metadata, the problems caused by semantic heterogeneity are described and the solution applied such as the use of controlled vocabularies, are discussed. Structured documents and the mark-up languages that they use are reviewed. The high cost of producing such documents means that production must either be sufficiently incentivised or be done by specialists. The rise of structured documents on the web (via XML) and the accompanying metadata standards such as PICS and Dublin Core are then covered. Finally, the uses of within-document metadata as a finding aid and in dynamic document assembly are discussed.

2.1 Definition

Most simply and most broadly, metadata is data about data. It has many uses, from library catalogues to non-textual corporate data (De Carvalho Moura, Campos & Barreto 1998). The research presented in this thesis, however, attempts to support goal-directed search and therefore is concerned with semantic document metadata, i.e. data about document content. Also, it is important that the metadata is human readable. Although some forms of metadata are meant primarily for consumption by computers, computers can only manage a relatively impoverished understanding of much content-descriptive metadata. Thus it is important that humans can directly use the metadata.

2.2 Types of metadata

Here we shall look at two different types of document metadata arising from two different disciplines: information science and artificial intelligence (AI).

The computerisation of library catalogues in the early seventies led to the development of sophisticated standards for bibliographical metadata. These are characterised by the Machine Readable Catalogue (MARC) (American Library Association 1996) and the more modern and flexible Encoding Archival Description project (EAD) (Swetland 1996). These standards define fields for a large range of information. A minority of these fields, such the 'subject headings' or 'subject' field, describe an item's semantic content. These 'subject heading' fields are often populated using standardised classifications systems such as the Library of Congress Subject Headings (LCSH 2002) (discussed further in section 2.3.1). Essentially, in this thesis, the metadata used is very similar to such subject headings, but instead of being applied at the document level, they are applied at the paragraph level.

In the artificial intelligence and knowledge-based system communities an ontology is most commonly defined as, "an explicit specification of [...] the relevant concepts of [...] some phenomenon in the world" (Benjamins et al. 1999, p. 691). It can contain a number of modelling primitives including classes, relations, functions, axioms and instances.

Ontologies can be used in the production of document metadata rich in formally encoded knowledge to enable the use of AI reasoning algorithms. They have been used to provide descriptions of news articles, academic literature and business data (Jokela, Turpeinen & Sulonen 2000; Motta, Buckingham Shum & Domingue 2002). In the business domain, TOVE provides a highly formal ontology for automated reasoning about business entities (Fox 1992). The enterprise ontology is less formal; its objective

is to enable the communication among people and computers in an organisation and assist in the "acquisition, representation and manipulation of enterprise knowledge" (Uschold et al. 1998, p. 3). Although such ontologies could be used to supply paragraph level metadata for business documents, they are pitched at too general a level for practical use; the enterprise ontology does not include terms for the many vital, yet idiosyncratic, organisation-specific elements.

At this point it is useful to distinguish between ontologies and taxonomies such as those developed in this thesis. A taxonomy is a strictly hierarchical classification of domain concepts exemplified by the Library of Congress Subject Headings (LCSH 2002). Ontologies range from the very formal, which include axioms and theorems about the relationships between concepts, to the informal, which are essentially taxonomies with the possible addition of non-hierarchical relationships.

2.3 Producing metadata

2.3.1 Managing semantic heterogeneity

Words and phrases mean different things, to different people, in different contexts (Furnas et al. 1987); this is the problem of semantic heterogeneity. In a corporate context the problem is exemplified by the differences in perspective, knowledge and language that tend to arise between different portions of organisations (Bryce 1997). How can metadata based descriptions of content deal with this problem?

One way of dealing with it is to have many different types of content description which together allow the reader to converge on a correct understanding. Both MARC and EAD do this. For example MARC has fields for summary (which allows abstracts), title, and subject headings. This solution helps the user to converge on the correct interpretation of *a* document's contents.

Another solution is the creation of what are called controlled vocabularies or classifications. Examples of this are the Medical Subject Headings (MESH 2002), and the Library of Congress Subject Headings (LCSH 2002). These systems provide a commonly agreed organisation of terms, so that large collections can be organised according to their content. This kind of solution has the problem of needing centralised control. Although this has been attempted (e.g. by the US government, Schuldt 1994), it would seem difficult to achieve in a corporate environment where knowledge is widely dispersed, not rigorously documented and things change rapidly. A potentially effective solution is to maintain a centralised glossary in a distributed fashion using hyperlinks from terms in documents. An example of this is the 'Virtual Hyperglossary' (West & Murray-Rust 1998).

An alternative approach is to encourage heterogeneity in the terms used as metadata while allowing selective and personalised views of the metadata. Huc et al. took this approach; data objects can be grouped into sets using unrestricted criteria based on metadata (Huc, Levoir & Nonon-Latapie 1997). Users of this proposed system are allowed to create their own hierarchies to describe the data from their perspective. Dourish et al have put forward a proposal which marries this approach with the use of controlled vocabularies. They allow standardised company taxonomies to be customised and adapted in such a way that mappings can be made between adaptations and back to the standard (Dourish, Lamping & Rodden 1999).

Hence, although semantic heterogeneity is a problematic aspect of the production and use of metadata, a number of approaches have been developed to deal with it.

2.3.2 Structured documents

The concept of a structured document can be likened to a form. The form explicitly dictates what goes where using mark-up languages, which use tags to attach metadata to data:

e.g. < wife's_name >Jane Bloggs</wife's_name>

This allows structure, in terms of what goes where or with what, to be predefined. It also allows metadata to be added to component parts of a document. Thus any document enriched with paragraph-level metadata will be a structured document.

2.3.2.1 Mark-up languages for structured documents and metadata

The mark-up languages used to create structured documents are generally described using either SGML (Standard Generalised Mark-up Language) (SGML:ISO 8879) or its younger net enabled sibling XML (Extensible Mark-up Language) (W3C 1998). These international standards allow mark-up languages to be flexibly defined in a device-independent manner that allows easy interchange of information. SGML requires what is called a DTD (Document Type Definition) to allow the tags to be defined along with the constraints on the kind of data they should contain, and their possible inter-relationships. In place of DTDs XML uses, but does not require, XML schemas.

SGML and XML have been used to create mark-up languages for textual documents. The Text Encoding Initiative (TEI) uses SGML for the detailed encoding of literary and linguistic textual materials in electronic form (TEI 2002). Encoding documents using the TEI DTD makes an extensive amount of information about the composition, style, and structure of a document available to computer analysis. Although TEI can be used to create paragraph level metadata, it is overly complex for this purpose, since it was primarily designed for detailed linguistic and stylistic analysis.

On the web XML has been used to define a number of mark up languages for encoding metadata. RDF (Resource Description Framework) is an XML standard for encoding metadata which can be used by AI reasoning algorithms (W3C 1999). Topic-maps are an ISO standard which allow metadata to be encoded so the metadata from different repositories can be easily combined (Pepper 1998). Although both of these standards ostensibly allow documents to be encoded with metadata, they are made complex by their aspirations to enable AI reasoning, and allow easy merging of repositories, respectively. This complexity makes them unnecessarily cumbersome and difficult to apply for simple cases such as the encoding of paragraph-level metadata.

Two standards have been developed specifically for metadata about web page content. PICS (Platform for Internet Content Selection), allows web pages to be labelled with information that would indicate content unsuitable for children (PICS 1996). Coming from the library cataloguing tradition the Dublin Core (Weibel, Grodby & Miller 1995) defines a list of 15 tags/fields for the description of resources. Two of these, 'subject' which contains keywords, and 'description; which contains a free text description, describe the resource's contents. Unfortunately, PICS and Dublin Core are geared towards document-level metadata, whereas here the metadata needs to be encoded at the paragraph-level.

This brief review suggests that the existing standards for document and metadata encoding are not ideal for the production of paragraph-level metadata undertaken in the research behind this thesis. However the basic mark-up languages, particularly XML, will be extremely useful in preparing documents encoded with paragraph-level metadata.

2.3.2.2 The pros and cons of structured documents

The use of structured documents has several benefits. Harris & Ingram (1995) report that at IBM using structured documentation of microelectronics allowed tailored paper documents to be provided very cheaply using constantly updated and reusable document components (Harris & Ingram 1995). Prettyman reports that converting a set of documents in the National Library of Medicine to SGML provided enough payoffs to make it cost effective (Prettyman 1996). Marked up documents can be published to tight deadlines simultaneously on paper and electronically with a consistent appearance in both. The SGML encoding also allows automatic collection of bibliographic information, and algorithmic conversion into new formats or forms.

These benefits do not arise from simply sprinkling a document with SGML. They come from the consistent encoding of rich content and structural information. This encoding takes time and effort. Hence, in order to maximise the cost to benefit ratio one must carefully decide what needs to be tagged. Mostly this kind of tagging is done by specialists, but if structured documentation is to be used on a very large scale, then non-specialists will have to play a large role, this however brings its own problems. Employing specialists to tag documents is not appropriate for many situations and is inherently expensive. Nevertheless, many attempts to encourage the producers of documents to do the tagging founder. The principal problem is that "the envisaged beneficiaries of the system simply do not have the motivation or time to invest [in tagging] to reach a critical mass of material" (Motta, Buckingham Shum & Domingue 2002, p. 2). Shum suggests that this problem can be overcome if the document users and producers have a culture of co-operation, the tagging delivers worthwhile benefits, and tagging is designed to fit in with existing work practices. Furthermore, it is possible to lower the cost of tagging, thus increasing the likelihood of it being done, by partially

automating the process using automatic classification algorithms (Joachims 1999). However it is clear that the issue of metadata authorship is a substantial barrier to the adoption of technologies which rely on information seeking technologies which use metadata.

Consequently, it is important to explore and warrant a range of metadata applications. If document metadata has a number of parallel uses (e.g. between document search, within-document search, dynamic document generation, etc) then its production will be more easily justified in terms of cost. If some of these uses of metadata deliver a visible and immediate payoff (e.g. within-document search and navigation) this could incentivise the involvement of non-specialists. Thus the development and evaluation of a new use for metadata could help motivate metadata production.

So although creating structured documents can deliver substantial benefits it is costly and it is difficult to involve non-specialists in production. The development of new valuable uses for structured documents and the metadata they contain could help justify their cost and incentivise involvement by non-specialists.

2.4 Using metadata

2.4.1 Metadata as a finding aid

Until the 1990's metadata was the principal means by which information retrieval was performed. Users searched library catalogues for documents using metadata field specified in standards such as MARC (American Library Association 1996). Such searching uses Boolean queries, which despite offering the potential for powerful search contain some inherent usability problems which have limited its take up by non-specialist users (Hearst 1999).

Initiatives, such as Cultural Heritage Information Online (CHIO) and the TEI, support Boolean queries on metadata for within-document searching. CHIO is a system with a text-encoding scheme developed for the museum industry in order to allow museum goers access to detailed multimedia information online to answer specific questions. The systems designers collected a large number of questions from their customers and used this to formulate a detailed schema specifying the kind of information of interest such as events, objects, people, places, and styles (Janney & Sledge 1995). The body of a document is marked up to indicate what each part is about (primary access points) and what is mentioned in passing e.g. people and places (secondary access points). Hence, user queries can be answered with the precise parts of the text that are of relevance. CHIO provides a good example of how metadata can be used at the within-document level to facilitate search by taking users directly to the information they need. The Text Encoding Initiatives (TEI 2002) as discussed in 2.3.2.1 enable those involved in literary study to make the many queries involved in their linguistic, stylistic, and textual analyses.

Highly structured metadata linked to formal ontologies is used for more sophisticated information retrieval technologies. One example of this is the KA^2 project (Benjamins et al. 1999). Here the home pages of researchers were augmented with metadata defining the meaning and type of information in each of a page's individual elements. Using this information and an ontology their inference engine could deliver precise answers to queries. For example, asked for all researchers in the KA^2 community, the inference engine not only retrieved those researchers who have annotated their pages, but also those with whom they have collaborated. This inference is made using the definitions of 'community' and 'collaboration' in the ontology. The W3C want the world wide web to become infused with this kind of metadata so that this kind of

intelligent search technology can be used on a large scale (Berners-Lee, Hendler & Lassila 2001).

2.4.2 Dynamic document creation - microdocuments and metadata

There are a number of systems which use SGML to locate and dynamically recombine sub-document chunks of text. These sub-document chunks of text are referred to as microdocuments. A microdocument is a meaningful chunk of text that might be of interest in isolation from the rest of the document it resides in. Microdocuments are invariably encoded using SGML or XML, governed by a DTD or XML-Schema. The majority of microdocument implementations focus on the ability to dynamically recombine microdocuments into new documents (e.g. Harris & Ingram 1995; Ahonen et al. 1997). In a sense paragraphs that are individually encoded with metadata can be seen as microdocuments although not necessarily ones that are intended for dynamic recombination into new documents. However this thesis is focused on how to support information seeking rather then document recombination. Ahonen et al use the document recombination approach to support information seeking. They built a system in which microdocuments relevant to a users query are collected into a single new document (Ahonen et al. 1997). There have also been attempts to use a similar approach to provide enhanced access to information by assembling documents in response to queries. This suggests that paragraph-level metadata could also be used to dynamically create documents or views of documents that are customised to the user's information needs.

2.5 Section summary

Two traditions in the production of metadata are described, library cataloguing and ontology based metadata in artificial intelligence. The metadata dealt with in this

research owes more to the library cataloguing tradition in that it is intended as a finding aid rather than as a basis for reasoning. The production of metadata, the problems associated with this and their solutions are then described. The problem of semantic heterogeneity can, in corporate environments, be tackled using controlled vocabularies. The production of metadata carries a high cost which means that it must either be sufficiently incentivised or be done by specialists. New standards for creating structured documents, such as XML, and producing metadata such as PICS and the Dublin Core, are reviewed. These lay solid technical ground for the creation of documents enriched with paragraph level metadata. A number of initiatives and research projects are reviewed which use within-document metadata as a finding aid. These illustrate that within-document metadata can be used to more directly satisfy information needs by: taking the user to the relevant text, directly answering the query, or dynamically assembling a new document containing relevant text from multiple documents.

3 Information visualisation

Card et al define information visualisation as:

> "The use of computer-supported, interactive visual representations of abstract data to amplify cognition." (Card, Mackinlay & Shneiderman 1999, p. 6)

Information visualisations can amplify cognition by: reducing the amount of cognition and working memory required for a task, leveraging the pattern detection and visual inference abilities of the human visual system, and by allowing exploration of the information space through interaction (op. cit.).

This review attempts to cover those parts of the information visualisation literature that are relevant to the design work carried out for this thesis. GridVis, the product of this design work, integrates and adapts a number of existing visualisation techniques. It visualises metadata housed in a relatively small taxonomy (with under 20 nodes) using a Windows Explorer style tree visualisation. The location in the document of a large number of metadata items is represented by combining a miniaturised image of the document with an adaptation of an existing visualisation named TileBars.

Thus this section, structured and unstructured layout techniques will be described and arguments will made for the advantages of structured layouts, such as the tree and grid layouts used in GridVis. Then, a number of the important interaction techniques from the information visualisation literature, which are employed in GridVis, will be described. Finally, the document visualisation literature will briefly be reviewed.

3.1 Layout techniques

Appropriate layout techniques are essential in any information visualisation. They determine the amount of information that can be displayed and the effectiveness with

which it can be read. There are different approaches to categorising the various layout techniques available; for example, Shneiderman's task by data type taxonomy (Shneiderman 1996). Here the techniques will be divided into those which are unstructured, and those which are structured, to emphasize the concerns important to the design decisions made in this thesis. In unstructured layouts; the data is used to determine the visual layout, in structured layouts the data is squeezed into a predetermined visual layout, such as a grid.

3.1.1 Unstructured layouts

Many of these techniques were pioneered in the seminal SemNet system (Fairchild, Poltrock & Furnas 1988). They attempted to visualise large knowledge bases by representing them as directed graphs in a 3D space. They used three dimensions because this allowed them to layout their graph without arcs between nodes intersecting each other, thus avoiding the visual confusion that intersections produce. They explored a variety of layout algorithms to make the structure of the knowledge base apparent.

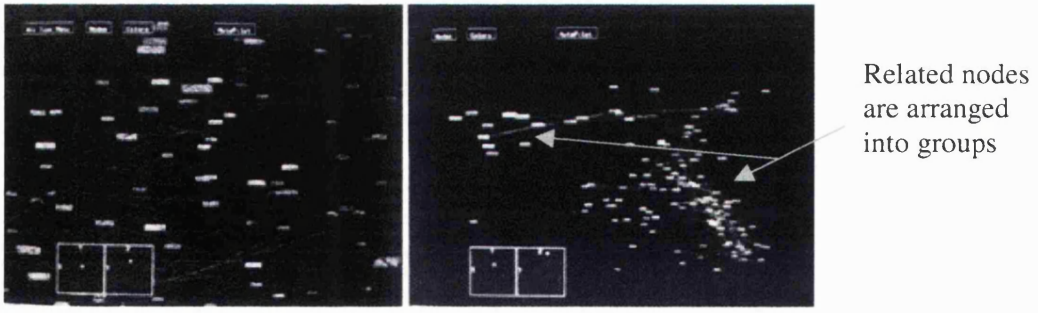


Figure 2 SemNet (Fairchild, Poltrock & Furnas 1988). The image on the left shows a network with random placement of nodes while the other shows it once the graph layout has been structured.

Multidimensional scaling and force directed placement were used to decrease the distance between nodes with large numbers of connections. A more sophisticated application of this idea using multidimensional scaling was embodied in Themescapes

(Wise et al. 1995). Hornbæk and Frøkjær compared this kind of interface to a simple boolean search interface (Hornbæk & Frøkjær 2001). They found that the visualisation did not increase the number or quality of documents retrieved, nor did it make searching faster.

Fisheye views, another technique experimented with in SemNet, also proved popular in information visualisation. This technique, invented by Furnas, was used to provide simplified views of portions of a graph (Furnas 1981). A fisheye view essentially displays details near a focal point and only more important landmarks further away (see 3.2.3 for more on this).

The great advantage of these unstructured layout algorithms is their ability to deal with large and complex data sets. They allow the structure in data sets to be seen. The draw back lies in the way that they display this structure. While many claims have been made for the simplifying power of these visualisations when a detailed evaluation was carried out, these advantages were not evident (Hornbæk & Frøkjær 1999). Consideration of these layouts from the perspective of the graphic design tradition (Mullet & Sano 1995; Tufte 1990) suggests that they are poorly equipped to support visual search and reference. Botsch and Kunz note that visual search is impeded by arbitrary non-aligned positioning of nodes, and that finding large patterns and following long paths is difficult and error prone (Botsch & Kunz 1996). In sum, the great draw back of these layout techniques is the lack of framing structure to help users to visually understand, orient in, and operate upon the data.

3.1.2 Structured layouts

Structured layouts represent data using predetermined visual structures, such as trees and grids. These structures can aid the user in navigating and understanding the data represented.

3.1.2.1 Tree structures

Tree structures[2] are a visual structure commonly used to represent ubiquitous (Robertson et al. 2002) hierarchical information structures such as file systems and taxonomies. They have also been used to represent portions of more complex structures, such as networks (Mukherjea, Foley & Hudson 2002), and intersecting hierarchies (Pollitt 1997; Robertson et al. 2002). Visual tree structures are problematic because, due to their generally high branching factor, they are often much wider than they are tall. This makes them difficult to fit on screen, and means they contain considerable empty, and thus wasted, space.

One strand of visualisation research has been concerned with fitting large tree structures on screen. The main problem here is that most common tree structures are much too broad to fit on screen in 2D and have text labels that are big enough to read. The Cone Tree solves this problem by wrapping child items round the base of a cone whose tip is the parent item (see Figure 3) (Robertson, Mackinlay & Card 1991). Items that are occluded can be brought into the foreground by rotating the cone. The Cone Tree visualisation has been used to represent the location of documents in a large medical taxonomy to provide an overview of the document collection and aid in query specification (Hearst & Karadi 1997). Another approach to showing an entire tree is to use an overview window, which shows the entire tree in miniature, and can be used to select a region of the tree for detailed inspection in a separate window (Korn & Shneiderman 1995). This visualisation will be described in more detail in section 3.2.2.

---

[2] A tree structure comprises items whose relationships branch out to child items which in turn branch out to their own children and so on, so that no relationships are directed up the tree towards ancestor items.

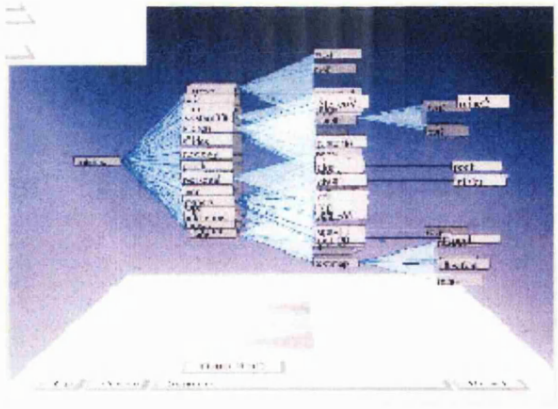


Figure 3 Cone Tree (Robertson, Mackinlay & Card 1991)

Alternatively, the problem of viewing large trees structures can be tackled by simply reducing the size of parts not being inspected. This approach was taken for the hyperbolic tree browser (Lamping & Rao 1996). This visualisation provided a smooth drop off in magnification from the centre of the screen towards the edges. The user can thus drag the interesting portions of the tree to the centre for inspection. Risden et al have developed XML3D, a 3D hyperbolic view of large tree structures coupled with 2D list panes showing the parents, siblings and children of the currently selected node (Risden et al. 2000).

Another strand of research has looked at representations which temporarily hide parts of a hierarchy. The cone tree allows the user to 'prune' and 'grow' branches of the tree using gestures, onscreen buttons, and keyboard short cuts (Robertson, Mackinlay & Card 1991). More recently, Robertson et al have developed a technique named visual pivot which only shows the common ancestors of selected items in huge intersecting hierarchies (Robertson et al. 2002).

Perhaps the most successful approach is the vertical indented folding tree popularised in file managers such as windows explorer. The combination of a left to right arrangement

of the tree and the ability to 'prune' and 'grow' branches, enables large trees to be explored in an interactively and computationally undemanding manner.

Due to its wide acceptance, the vertical indented folding tree tends to be used as the control in the evaluation of the more exotic tree visualisations. In general, it proves to be the most effective visualisation for most situations. The cone tree resulted in slower search times which deteriorated more rapidly as the branching factor of the data structure increased (Cockburn & McKenzie 2000). The hyperbolic tree produced search times that were indistinguishable from Microsoft's Windows Explorer for wide range of search tasks (Van Der Wege & Pirolli 2000). However Windows Explorer was found to be consistently slower, for the task of finding categories in an extremely large (12,000) category hierarchy, than XML3D, the 3D hyperbolic browser described above (Risden et al. 2000). Although the XML3D visualisation seems well suited to very large hierarchies, it is not clear whether its advantage would persist with smaller trees. Taken together this evidence suggests that the humble vertical indented folding tree visualisation stands out as one of the best performing tree visualisations for all but extremely large hierarchies.

In summary, this section has briefly covered the different approaches to using tree structures in visualisation and argued that the vertical indented folding tree has proved itself the most successful representation for a wide ranges of purposes.

3.1.2.2 Grid structures

The grid is another powerful visual structure that can be used in visualisations. Grids have been exploited extensively in graphic design (Mullet & Sano 1995) and in interface design (e.g. spreadsheets), but less so in information visualisation. However,

grids have been used in some information visualisations; a selection of these are reviewed here to illustrate techniques used in this thesis.

As part of a tool to encourage the use of online learning resources, Ding and Marchionini use a grid structure to provide an overview of the available teaching resources (see Figure 4) (Ding & Marchionini 1998). The rows and columns can be assigned to attributes such as topic and data type. The glyphs in the cells of the matrix indicate the relative number of resources either by the number of dots in cell, or when there are more than 50, by the length of a bar. The type of resource is represented by the colour of the glyph. A problem with this visualisation is that it is very limited in the size and complexity of attribute set it can use. In addition, the use of both dots and bars to represent magnitude results in the misleading impression that the cells with only bars in them have fewer items. However, it does illustrate how the use of glyphs in the cells of a grid can communicate category (by colour) and magnitude (by number or size).

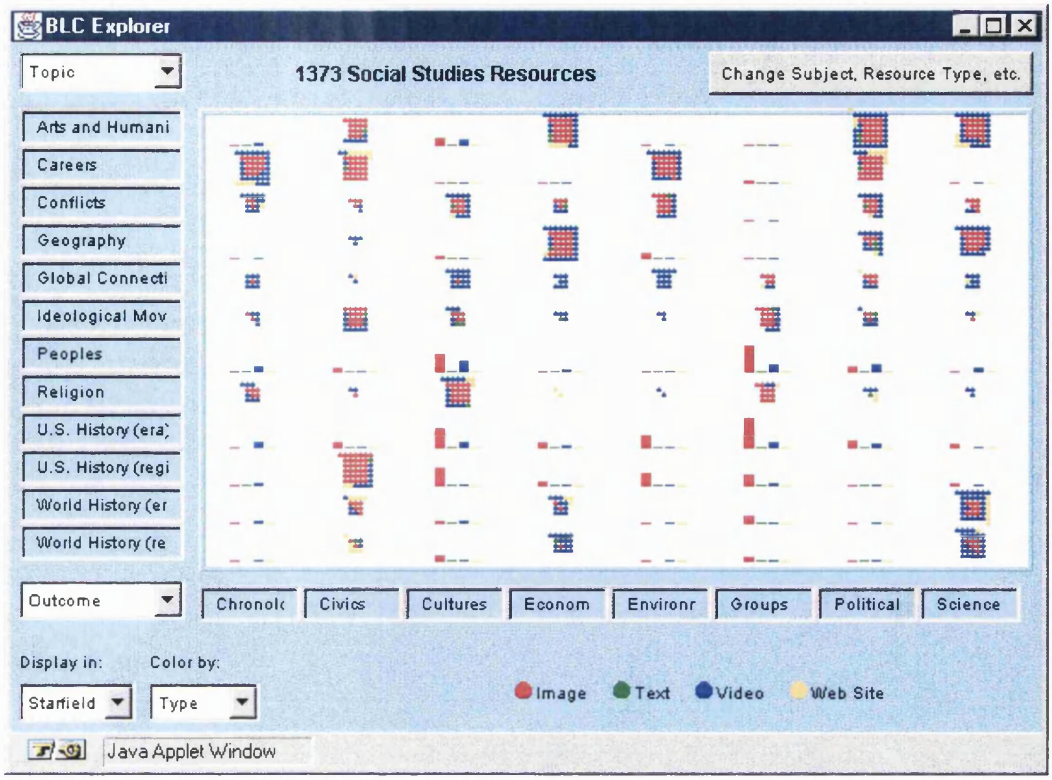


Figure 4 Overview of teaching resources (Marchionini & Shneiderman 1988)

Another visualisation, the Matrix Browser, combines the grid structure with a vertical indented folding tree to allow the visualisation of large networks (Botsch & Kunz 1996)(see Figure 5). Hierarchical structures, abstracted from the network, can be displayed on the vertical and horizontal axes using the familiar vertical indented folding tree. Relationships between nodes are revealed by lines extending out orthogonally from the trees into the grid. Different aspects of the network can be examined by alternating between different hierarchies. The use of the folding tree means that the grid can be used for a large number of items.

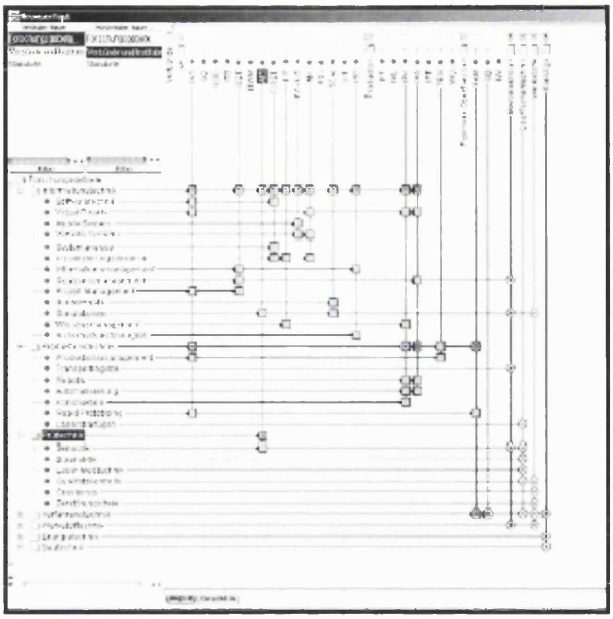


Figure 5 The Matrix Browser (Botsch & Kunz 1996)

This section has reviewed the different uses and embodiments of grid structures in information visualisation. It shows how information can be displayed in cells using the colour, number and size of glyphs. It also illustrates how the number of items that can be displayed in the grid is increased by the integration of a vertical indented folding tree. Although there appears to be no empirical work which examines the benefits of grid representations, Botsch and Kunz claim that, relative to unstructured representations, they ease visual search (Botsch & Kunz 1996). Sano and Mullet argue that they allow users to benefit from experience as they learn to predict where information will be found (Mullet & Sano 1995). In contrast, Hearst notes that grids are limited to showing the intersection of two attributes and since many of the cells are often empty they can be space-inefficient (Hearst 1999).

3.2 *Interaction techniques*

Having looked at available layout techniques in section 3.1, the various interaction techniques used in the design work presented in this thesis are now introduced and their advantages are evaluated.

3.2.1 Dynamic querying

The development of dynamic querying was an attempt to improve on the standard database search interfaces. In a dynamic querying interface the results are presented visually and the database is searched by visual manipulation of the query components. The results are rapidly (<100ms) displayed so as to give the user reversible, incremental control of a query (Shneiderman 1994). When compared to natural language and paper based interfaces, users answered queries faster with, and gave higher subjective ratings to, a dynamic query interface (Shneiderman 1994). This kind of interface lets users interactively explore multidimensional databases and discover trends, clusters, gaps and outliers.

3.2.2 Tightly coupled displays

This ubiquitous technique involves the dynamic linkage of multiple views of the same data. Perhaps the oldest application of this idea is in scatterplot *brushing* (Becker & Cleveland 1987); points on one scatter plot (say film revenues vs. date) are selected and the sampled data points are highlighted on another, related scatterplot (say rainfall vs. date). Brushing allows the user to explore multiple facets of particular portions of large data sets.

Another important application of this technique is in the provision of overview maps which link to a window providing a detailed view, a technique called *overview+detail.* The PDQ Tree-browser (Pruning with Dynamic Queries) uses coupled displays to

provide access to large tree structures through an overview window showing the entire tree, and a detail window in which the node labels can be read (see Figure 6) (Kumar, Plaisant & Shneiderman 1997). Dynamic queries can then be constructed which grey out the unwanted portions of the tree in both views. A wire frame box in the overview window, the *panner*, determines what is seen in the detail window. So, if all the material currently in the detail view has been greyed out, the panner can be moved to bring the areas not greyed out by the dynamic query into view.

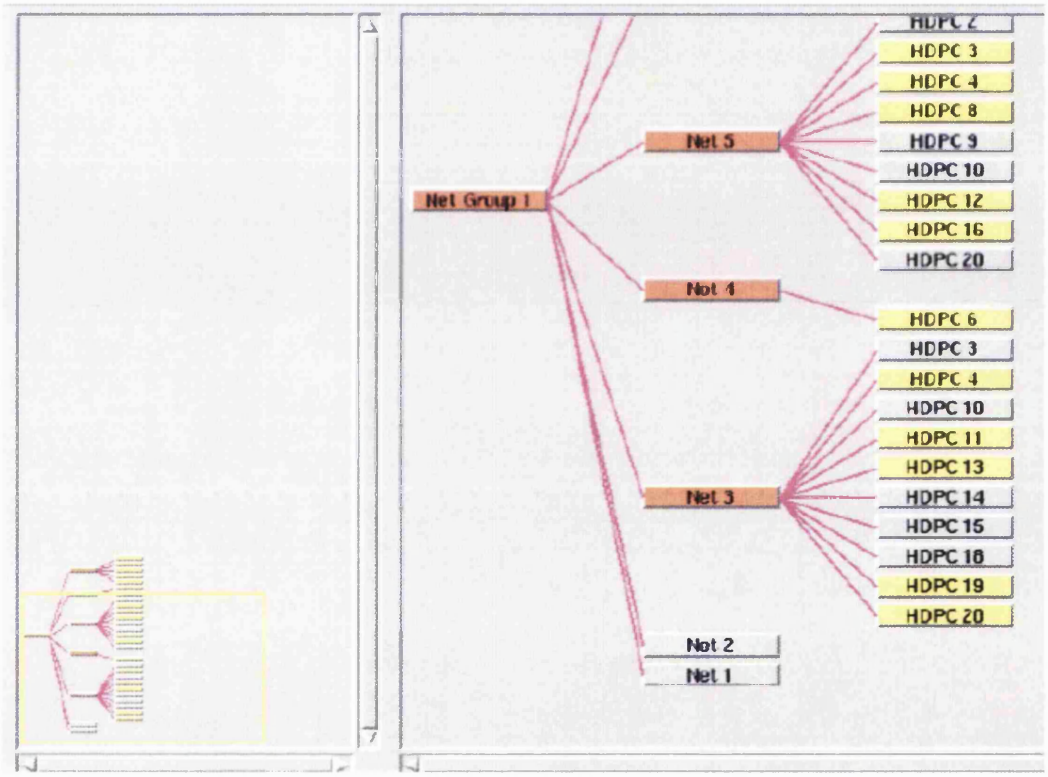


Figure 6 PDQ Tree-browser (Kumar, Plaisant & Shneiderman 1997)

North and Shneiderman report an empirical examination of the effect of coupling overview+detail displays of a large document (North & Shneiderman 2000) (see 3.2.3 for more on overview+detail). There were three conditions: a single 'detail' view with the entire document in a scrolling window, the detail view with an overview window showing the section titles, and detail and overview widows coupled so that selecting a

heading in the overview window scrolls the detail window to the corresponding section. The experimenters found that the uncoupled overview was only faster on tasks that required participants to find overview information, but when coupled with the detail window it proved to be on average 80% faster for search tasks. This shows that the combination of tight coupling and overview+detail displays can deliver substantial improvements in navigation performance.

Figure 7 The fisheye menu (Bederson 2000)

3.2.3 Focus + context displays

Many methods have been developed to provide single views which simultaneously preserve overview while giving access to detail.

Spence and Apperly developed a focus+context visualisation, named the bifocal lens, for using and locating information in the workplace (Spence & Apperly 1982). Furnas generalised this technique calling it a fisheye view (Furnas 1986). This method is based on the creation of a level of interest metric and manipulation of the objects in the display to emphasise objects which score highly. The degree of interest metric is calculated as the sum of an a priori measure of importance and the distance from the point of interest. For indented program listings, importance is based on how indented the text is and, the distance, is calculated as the number of lines away from the point of interest. A simple algorithm can be used to omit lines with a low degree of interest, preserving all the code

at the point of interest, but only the programs skeleton on either side of this point (op. cit.). This principle has been applied in many visualisations (Leung & Apperley 1994). Recently, the fisheye technique was applied to the pull-down menu widget and provided satisfying and rapid access to items in long menus (Bederson 2000) (see Figure 7)

3.3 Corpus visualisation

Much work has been done on techniques for the visualisation of large document collections. Much of the literature investigates techniques for visualising the outputs of content analysis, which are sophisticated analyses of term frequency statistics (Carey, Kriwaczek & Ruger 2000; Wise et al. 1995). This section will, however, concentrate on those two aspects of this literature that are relevant to the current research: the visualisation of document metadata in general, and the visualisation of taxonomies.

3.3.1 Visualisation of document metadata

Beagle argues that the use of metadata for document visualisation, as opposed to full text, would increase the economic viability and end-user accessibility of visualisation technologies (Beagle 1999). Firstly, content analysis of full text documents is very demanding computationally, whereas metadata by its nature would be much faster to process. Secondly, metadata is easier to transfer and manipulate due to its smaller size, and thirdly, it is not subject to intellectual property rights restrictions.

Despite these attractions, there is currently little visualisation work which concentrates on visualisation of metadata; one exception, however, is the visualisation of co-citation. Chen and Paul use a number of mathematical techniques to extract the strongest links between authors and lay the resulting simplified network out in a plane (Chen & Paul 2001). The network is then coloured according to the primary specialities of each author as calculated by factor analysis and multidimensional scaling.

The current work on the creation and use of RDF and Dublin Core metadata for the web is likely to result in a large increase in the amount of useful metadata available. This could open the way for more visualisation work based on document metadata.

3.3.2 Visualisation of category hierarchy

The many category hierarchies used to organise document collections have motivated research on category hierarchy visualisations that are tightly integrated into the search process. The majority of the innovations are in response to the great size and complexity of the category hierarchies being visualised. For example, the Medical Subject headings (MeSH) tree has approximately 18,000 terms, and contains non-mutually exclusive high-level categorisation systems such as grouping by 'body region', and 'biological system' (Korn & Shneiderman 1995). The work in this thesis however deals with small taxonomies. As such, the emphasis will be placed on the support given by these visualisations to information seeking.

MeSHBROWSE diplays subsets of the MeSH hierarchy as a 2-D horizontal tree (Korn & Shneiderman 1995). This allows user to navigate through the category hierarchy, as well as to explore hidden inter-relationships among categories. Hidden relationships, e.g. items in the 'biological system' part of the tree, which exist in the same 'body region' are shown by colour highlighting of items related to those selected. Multiple selections are equivalent to a Boolean AND query, highlighting just those items which relate to all of the selected items.

The Cat-a-Cone system, from Xerox PARC, also visualises the MeSH Hierarchy, but it offers closer integration with the search process (Hearst & Karadi 1997). The use of the Cone Tree visualisation (described in section 3.1.2.1) allows large portions of the category hierarchy to be displayed. Users can navigate through the hierarchy using the visualisation, or they can undertake text searches of both the category labels and the

document text. When inspecting the results of a free text search, the Cone Tree reveals the parent categories of each result. This allows users to spot otherwise obscure, yet useful, categories and can be helpful in interpreting category labels.

These two systems illustrate that very large category hierarchies can be visualised to allow them to play a useful part in the retrieval process. They also illustrate the interactive techniques that can be useful with category hierarchies. Two of these are adopted in the design work presented in this thesis: selection and highlighting to reveal relationships between categories, and automatic display of relevant categories (and their ancestors) upon selection of documents.

3.4 Single document visualisation

As stated in chapter one there has been relatively little work done to improve the digital access to information within documents. This section reviews the three threads of existing research which relate to this problem. The first visualises the distribution of query keywords within documents to either provide richer search results or to support navigation. The second thread describes general visualisations of document content for navigation. The third thread concerns the visualisation of narrative themes within a document.

3.4.1 Visualisation of key word distribution in document surrogates

Some visualisations use a display of distribution and prevalence of query terms primarily to inform relevance judgments and secondarily as a navigational affordance. This area is particularly relevant here since visualising the distribution of query words in a document is somewhat analogous to visualising the distribution of relevant paragraph-level metadata. The different visualisations of keyword distribution will therefore be described so that useful techniques can be drawn out.

The first of these is Hearst's TileBars visualisation (Hearst 1995) (see Figure 8). She visualized the relevance of search terms to different parts of a document. Documents were divided up into topics and the relevance of each to a query was represented by a square: the darker the square the higher the relevance of that topic. The squares were arranged in a row, or bar, to represent the entire document. Several separate term sets can be incorporated at once by visualising the relevance of each with a separate bar.

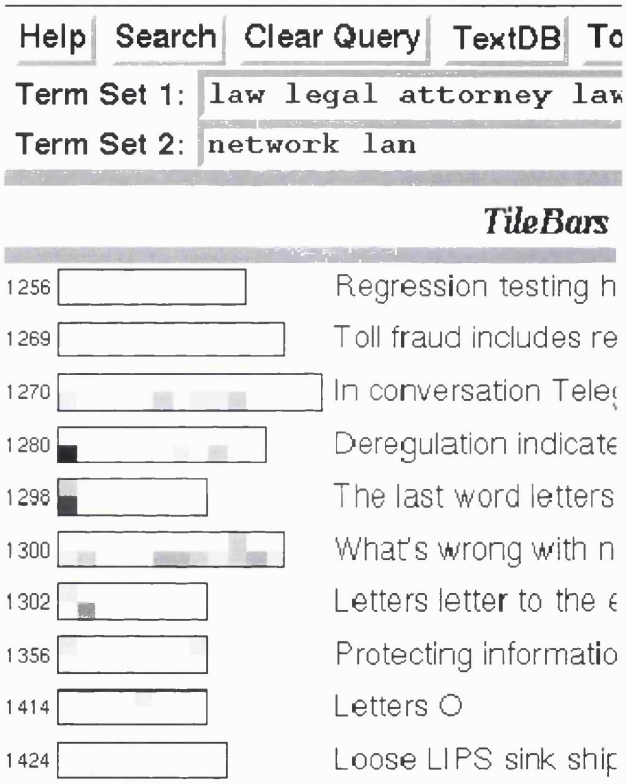


Figure 8 TileBars (Hearst 1995)

This visualisation has several advantages: it is simple and compact enough to be tightly integrated within a traditional search result display and it shows the degree of relevance and query term co-occurrence as well as relative document length.

Unfortunately, TileBars was not found to offer a quantitative improvement in the accuracy of relevance judgements (Ogden, Davis & Rice 1999). However, TileBars can also be used to navigate to the most relevant part of the document. Selecting a particular cell within a tile bar will open up a window, with the document scrolled to the selected point. The design developed in this thesis borrows from the TileBars design, although it is to support navigation and search, rather than relevance judgement.

Visualisation of query keyword prevalence has also been expressly used to support navigation to relevant portions of a document. The first system to experiment with this was SuperBook (Egan et al. 1989). When a keyword search was performed each entry in the fisheye indented tree visualisation of the document's table of contents was annotated with the number of times the keyword appeared. Sections containing the keyword were expanded to reveal the relevant subsections while irrelevant sections were folded away leaving only high-level chapter headings. This allowed the user to identify the sections of the document most relevant to the query 25% faster and more accurately then with a paper version of the document (Egan et al. 1989). Because this approach to visualising content distribution relies on headings it is particularly suited to large documents that are divided up into fairly small sub-sections.

The Reader's Helper makes a scrollbar out of a document overview in order to support the reading process for shorter documents (Graham 1999). It consists of a web browser with a thumbnail of the entire web page alongside it. This can be used as a scroll bar by vertically dragging a rectangle representing the document portion currently in view in the browser. Relevant keywords and phrases are highlighted in both the thumbnail and the overview. These keywords and phrases relevant to the reader's interests are identified via a user profile and probability-based content recognition. The different 'concepts' defined in the reader profile are listed to the right of the document and can be switched on or off. The system also provides the option to print a view of the document which includes the keyword highlighting, as well as a one-page document overview using thumbnails to depict each printed page.

Suh et al make a number of criticisms of this design (Suh et al. 2002). They point out that the user profiles are an application setting and can thus not be selected on the fly which limits the kinds of reading task for which the design is useful. The use of only

one colour for highlighting means that the co-occurrence of different concepts cannot be investigated, and the use of different colours in the thumbnail and detailed view make the transitions between views harder.

Suh et al put forward their PopOut Prism (see Figure 9), which can be seen as a refinement of the visualisation design in the Reader's Helper system (Suh et al. 2002). They provide a similar interface where the search terms entered appear as colour labels indicating the location of the keyword in both the browser window and thumbnail. The thumbnails and labels have been carefully designed to be legible in both the overview and document, to draw attention to the labels, and to minimise the cognitive processing required to make the transition between views.

Figure 9 The PopOut Prism (Suh et al. 2002)

A qualitative study suggested that the PopOut Prism supported visual search strategies in three ways (Suh et al. 2002). First, users used the density and number of colour labels to identify particularly relevant portions of a page. Second, they ignored some clusters of labels due to their location, for example when they were over the menu bar. Third, the spatial configuration of labels proved useful; for example, users would look for clusters of labels with different keywords, to identify particularly relevant sections of the document.

This section has described research on visualising the distribution of keywords in a document. Although there has been little quantitative evaluation of the utility of keyword distribution information itself, qualitative evaluations indicate that users regard it as useful (Byrd 1999; Egan et al. 1989; Suh et al. 2002). A possible problem with these interfaces is the fallibility of keyword data, for example, keywords in the text might not have been used in the sense being sought. Nevertheless, the distribution of relevant key words gives a first indication of the parts of a document that are relevant to a query.

A number of design guidelines can be drawn from this body of work. Firstly, a visualization for supporting goal-directed search should use an overview+detail paradigm; using the overview to convey the location and relative number of relevant topics. Secondly, this overview should double as a tool for navigation. Thirdly, when using multiple search words, it should be possible to distinguish their individual distributions.

3.4.2 Visual document navigation and querying systems

A number of visualization systems have attempted to facilitate document navigation by displaying the entire document on screen.

The first of these is the Document Lens (Mackinlay & Robertson 1993) (see Figure 10), which allows the entire document to be viewed at once, fully displaying the page of interest while keeping the rest of the document viewable, if not readable. Keyword searches can be performed and the results are displayed by highlighting the word wherever it appears in the document. The mode of overview here is primarily visual as opposed to semantic, since text outside the focus region is very hard to read. It is unclear whether this kind of overview is very helpful and there has been no published evaluation to suggest that it is.

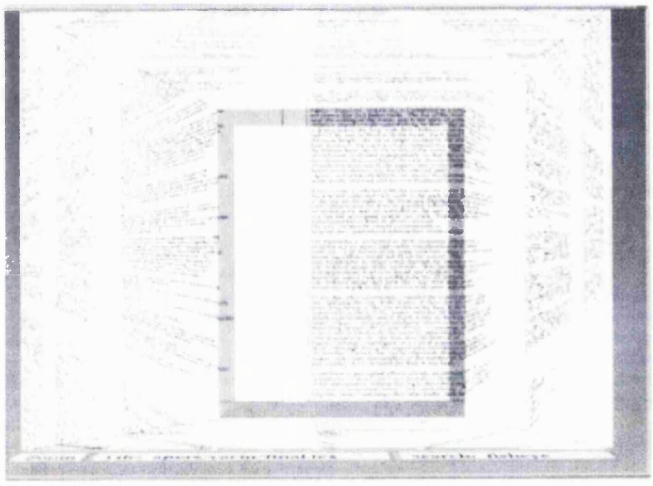


Figure 10 The Document Lens (Mackinlay & Robertson 1993)

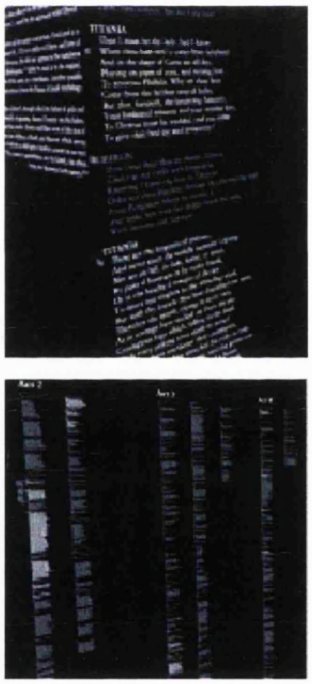


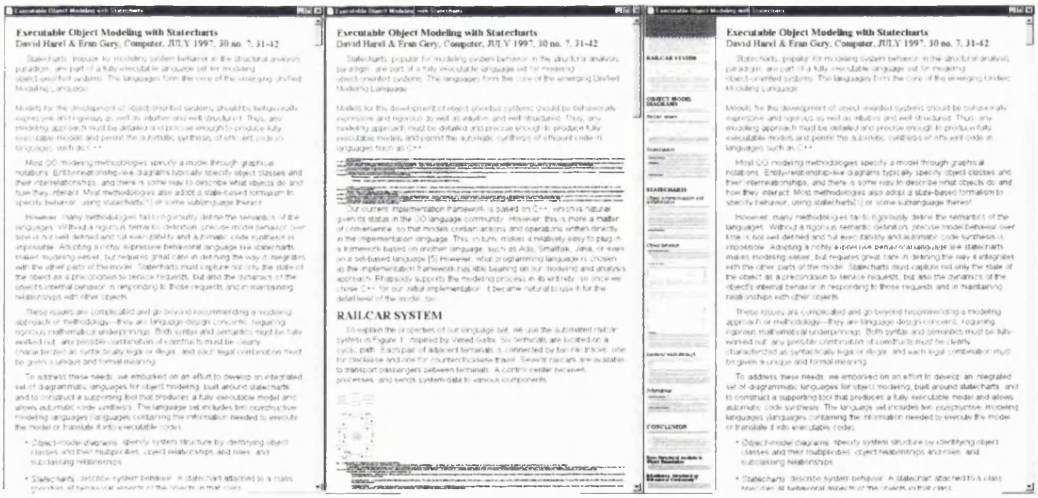
Figure 11 Text in 3D (Small 1999)

Small experimented with alternative ways of laying text out in three dimensions (Small 1999). The plays of Shakespeare were laid out so that the structure of the play could be seen all at once (see Figure 11). This kind of design only makes use of the third dimension as a metaphor to support zooming. The feature most useful for goal directed reading, the highlighting of portions of text relevant to queries, can easily be handled with a 2D overview of the text (see 3.4.1). Also, any advantages available for goal-directed reading would need to outweigh the difficulties introduced by the use of 3D, such as the difficulty of navigating in 3D space and of reading text that is not parallel to the view plane (Small 1999).

Another 3D visualisation of single documents is the 3D book (Card, Robertson & York 1996; Woodruff et al. 2000). Here an entire book or a collection of web pages is represented using a 3D book metaphor. The user is able to turn to specific pages, turn

pages one at a time, or rifle through the pages seeing what is on each page as it turns. This visualization offers very familiar navigation and a good metaphor for presenting collections of information. Nevertheless, it does not directly attempt to support goal-directed search, and it is not clear that it would be able to do so.

Turning to 2D interfaces, Hornbæk and Frøkjær applied two standard visualisation techniques to the support of reading and evaluated their effectiveness (Hornbæk & Frøkjær 2001). They compared a standard linear document browser to a overview+detail interface and a fisheye interface. The three interfaces used in the evaluation are shown in Figure 12. In the fisheye interface, traditionally more informative portions of the document (such as the first and last paragraphs) are at a readable size, the rest are much reduced – the user can expand these by clicking on them. The overview+detail view is much like those described in 3.4.1, except that headings are maintained at a readable size.



a) Linear b) Fisheye c) Overview+Detail

Figure 12 Three document visualisations (Hornbæk & Frøkjær 2001)

Two tasks were evaluated, one was characterised as 'reading-to-learn-to-do', and the other was a goal-directed reading task. In the 'reading-to-learn-to-do' task, the

interfaces were used to read a document. Performance was judged by measuring reading times, and scoring subsequently written précis and answers to unanticipated questions. In the goal-directed type of reading, performance was measured by giving the participants six questions, one at a time, recording task times and marking their answers.

The results of this analysis favoured the overview+detail interface. It produced significantly higher subjective satisfaction scores. It also delivered significantly better answers for the 'reading-to-learn-to-do' task, although not for the goal-directed one. However, the overview+detail interface resulted in significantly slower performance in both tasks. The fisheye interface resulted in worse question answering and subjective satisfaction scores than the linear and overview+detail interfaces, although it was faster for the 'reading-to-learn-to-do' tasks. This evaluation suggests that the overview+detail interface visualisations allow for a better understanding of the document, although they can slow performance.

A number of design guidelines can also be distilled from the work reviewed in this section. Although several interfaces have used 3D, it is not clear that these would offer particular advantages for goal-directed search; additionally, no evaluations have been published to suggest otherwise. Therefore, given the additional complexity introduced by the use of 3D, 2D interfaces are currently preferable. Turning to 2D interfaces, an evaluation is described which shows the relative advantages, for supporting goal-directed reading, of two visualisation techniques, overview+detail and fisheye. This evaluation suggests that an overview+detail interface with a document thumbnail as an overview can improve understanding at the cost of slowing performance.

3.4.3 Visualising document narratives

Some researchers have used information from content analysis to extract and visualise the themes as they evolve through a document. A document's features such as term

frequency are used to automatically determine changes in the topic being discussed. Keywords which describe each topic can also be extracted, which, together with the data on the change in themes, can be used to visualise a documents narrative.

Perhaps the first work to emerge on this is TileBars, already discussed in section 3.4.1. Hearst and Plaunt devised a topic segmentation algorithm to automatically detect topic boundaries and extract keywords that were representative of each topic (Hearst & Plaunt 1993). However, her visualisation is designed to help assess search results rather than visualise the document's narrative.

Researchers at the Pacific Northwest National Laboratory (PNNL) have devised a text analysis and visualisation system based on wavelets (Miller et al. 1998). Textual features of a document are used to construct an input to an algorithm whose output, the wavelet energy, expresses the rate of change in narrative flow at each point in the document and at several levels of detail. They present this information to the user in several ways. The simplest is a set of charts, one for each level of detail, with the wavelet energy on the y-axis and document position on the x-axis. Another shows the entire document and maps energy levels to font colours in a miniaturised thumbnail view of the document. This analysis technique is effective at detecting narrative changes. Nevertheless, it is not clear which, if any, of these visualisations would actually aid the reading process because the visualisations are not embedded in a reading or a work process.

Brand expressly set out to visualise narrative themes (Brand 1997). He clusters words using a thesaurus and a maximum likelihood algorithm to produce pools of words with similar meaning. The highest ranked clusters are used to indicate thematic motifs of the story. These are visualised using the techniques developed by Small (see 3.4.2) (Small 1999). The full document text is laid out vertically in the background, and the keywords

describing each narrative theme are laid out in the foreground. Selecting a theme highlights the associated sentences. Unlike the PNNL work, this visualisation could easily fit into common reading practices. The user could open a document, read the theme summaries, select the one of interest, and zoom in to read that part of the text.

This work demonstrates the possibility of using content analysis to provide data for a document overview visualisation. However, more work is needed to establish that this is indeed an effective approach relative to alternatives described in sections 3.4.1 and 3.4.2.

3.5 Section summary

Unstructured visualisations for depicting large graphs have briefly been reviewed and it has been noted that the lack of regular structure may hinder visual search and reference. Some of the many visualisations based on tree structures were then reviewed. It is concluded that the vertical indented folding tree is, for most tasks, more effective than its more exotic counterparts. The use of grid structures in visualisation was then reviewed. A number of interaction techniques developed for use with visualisations were described. Dynamic querying allows rapid exploration of databases, the use of multiple views is simplified by tightly coupling the displays, and focus+context displays can aid navigation by giving access to detail as well contextual information.

The document visualisation literature was then examined. The potential advantages and relative paucity of visualisations of document metadata was highlighted. However, a number of useful techniques were drawn from visualisations of category hierarchies. In contrast, much work has been done on the related area of visualising keyword distributions. This research was reviewed, suggesting that overview+detail interfaces can be used to show the location of, and aid navigation to, relevant content. Work on

providing visual overviews of documents was then reviewed. This suggests that reduced thumbnail images of documents can lead to improved understanding of a document. A number of 3D interfaces for document navigation, are also reviewed. None of these interfaces offer compelling support for goal-directed search, and no evaluations have been published to suggest that they do. Finally, work which has attempted to visualise document narratives was reviewed; it is noted that evaluation work is needed to assess the effectiveness of this approach for concrete tasks such as goal-directed search.

4 Information seeking and information retrieval

This section briefly covers pertinent portions of information seeking and information retrieval literature. It draws on theories of information seeking to argue for the potential utility of the metadata taxonomies used in GridVis. Then it looks at document use in the workplace arguing that goal-directed search within-documents is prevalent and thus worth supporting. Finally, the continuing pre-eminence of paper as the medium for document use is noted and lessons are drawn for the design of tools for digital document use. It is also suggested that paper would thus be a suitable benchmark against which to evaluate the performance of GridVis.

4.1 Models of information seeking

Much research and theorising has been directed at building a detailed understanding of the information seeking process. Here the basics of the resulting theories are outlined, and the implications they have for the design of systems that support within-document search are discussed.

Hearst presents a simple information retrieval model which captures the basic structure of many different models in a simple general model (Hearst 1999). In this model there are two steps which merit further discussion due to the problems they can present to the user (Kuhlthau & Tama 2001; Nicholas & Martin 1993): these are the formulation of an, 'information need' and 'query'.

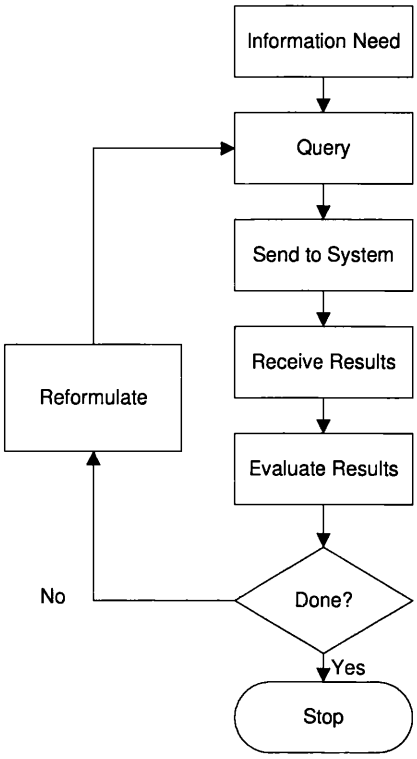


Figure 13 A general model of information seeking (Hearst 1999)

Taylor describes the process of forming an information need and articulating it as a query using four points or stages (Taylor 1968). An information need starts as an actual but unexpressed *visceral* need, this evolves into a *conscious* need, which can then be expressed, as a *formalized* need, but must ultimately be articulated through the constraints of the system as a *compromised* need. This process can be problematic since it involves uncertainty, which itself can lead to anxiety (Kuhlthau 1993). Progressing from a visceral to formalised need involves uncertainty "of a conceptual nature (i.e. what the user is looking for)" (Kuhlthau 1993 p. 343) and progressing from a formalized need to a compromised need involves uncertainty about "interactive technical process" (Kuhlthau 1993 p. 343) such as how to use the interface and what language to use. Search tools should therefore seek to support users by reducing the uncertainty inherent in the formulation of information needs.

The use of metadata as the basis for within-document visualisation might reduce this uncertainty in a number of ways. Metadata can be used to present the user with a manageable description of what information is available, thus serving as a prompt that can formalize an initially vague need. Similarly, the use of a limited and uniform controlled vocabulary to describe content may reduce uncertainty about how to translate the formalised need into a compromised one (Kuhlthau & Tama 2001).

4.2 The limitations of information retrieval techniques

Although information retrieval techniques have been hugely successful for handling large collections, they have limitations which make them less effective for use with single documents.

All widely used information retrieval techniques rely on term frequency data and string matching. Essentially words in the query are matched to those in an index of the text collection being searched. The index records the location for each word occurrence. Most simply documents with words that match the keywords are identified and ranked according to the number of matches they contain. More sophisticated algebraic, probabilistic and set theoretic approaches allow better use to be made of the index's term frequency data, resulting in more inclusive and more accurately ranked result sets (Baeza-Yates & Ribeiro-Neto 1999). Despite their sophistication, these techniques are all limited to working with terms rather than meaning, and so have a number of shortcomings.

The reliance on term frequency information means that relevant information will be missed due to the variation in vocabulary use and the sophistications of natural language. Blair and Maron found that actual recall, for full text searches of a large legal document collection, was much lower than its users expected (Blair & Maron 1985).

They found that this was partly due to variability in vocabulary, viz. the query words were related but different to those used in the documents. A series of studies by Furnas et al support this finding. They found that the probability that different people would use the same term for an object was less than 0.2 (Furnas et al. 1987). Blair and Maron also found that relevant documents were missed by search professionals due to the oblique use of references and treatments of a topic from a different point of view, which skirt around the query terms. These are important problems which come down to the limited 'understanding' that can be drawn from term frequency statistics.

Information retrieval and natural language researchers have made many attempts to overcome these problems, but have met with limited success. One approach is to use a thesaurus to suggest or add related terms to the query. Unfortunately this reduces precision and swells the result set. Latent semantic indexing is an alternative approach which uses term frequency statistics and a data reduction technique to map the query and documents to a low dimensional space associated with concepts (Deerwester et al. 1990). Documents close to the query are returned and ranked in order of increasing distance. This technique has generated considerable interest, although in practice it has yet to demonstrate its superiority over conventional techniques when used with general document collections (Baeza-Yates & Ribeiro-Neto 1999). A very different approach is to exploit syntactic information in order to extract more meaning from documents. However despite over 30 years of research, natural language processing techniques have not significantly improved document retrieval (Harman, Schauble & Smeaton 1997; Sparck Jones 1995). In sum, the various approaches taken to overcome the limitations of techniques term frequency based techniques have not met with much success.

It is important to note that for within-document search the failings of approaches which rely on term frequency statistics become more serious. The small number of words in

within-document spans of text, such as paragraphs, reduces the effectiveness of term frequency based techniques (Ingwersen 1996). Hence, techniques intended to improve problematically low recall, such as latent semantic indexing, will be even less effective for within-document search. However, while the huge amount of text in document collections means that there is no practical alternative to using some form of string search, given the small amount of text in single documents it should be possible to do better with them.

String searching itself has a number of other drawbacks. Although string searches can be an effective way of locating information, they do not show the user what information a document contains. Thus, the search tool does not indicate what type of information needs the document might answer. In addition, the requirement for the user to compose a query can, as argued above, make it difficult for the user to formalise their unexpressed 'visceral' information need.

Thus although string searching has a proven track record with document collections, other techniques might surpass it with single documents. The use of paragraph level metadata could, for relatively short single documents, avoid problems of low recall through the use of corporate taxonomies and the involvement of human authors. This metadata could also be used to display the contents of documents and act as a prompt for the formalisation of information needs.

4.3 *Information seeking, document use and goal directed search in the workplace*

A number of studies have examined the variety of information seeking and document use practices that occur in the workplace. These, together with other related research,

are used to argue that goal-directed search within-documents is prevalent in the workplace.

4.3.1 Goal-directed search within-documents

Document use is involved in a large proportion of the working day across professions. Adler et al calculate that this proportion is, on average of 82% of the working week across their sample, ranging from 23% for a nurse to 94% for an accountancy assistant. At the International Monetary Fund Sellen and Harper found that 97% of time involved working with documents (Sellen & Harper 1997). This suggests that tools to support document use could play an important role in the workplace.

The next question then is, to what extent does work place document use involve goal directed-search? Levy argues that there is a general trend towards what he calls 'hyper-extensive' reading which is the search and consumption of "just the right, the most relevant, minimal units or fragment" (Levy 1997, p. 208), or in other words goal-directed search within-documents. He goes on to argue that this is particularly true for on-screen reading which on the one hand provides search facilities and on the other hand has disadvantages for prolonged reading (the reasons for which are discussed below in section 4.3.2). Byrne et al report a study of web use, which provides some support to Levy's vision (Byrne et al. 1999). Over half of their video diary footage of every-day web usage comprised goal-directed within-document searching. This suggests that for online document use, at least, goal-directed search within-documents is highly prevalent. However, although Byrne et al took care to collect data from real-world web use, they did not concentrate exclusively on document use in the workplace.

Adler et al. however do report a large diary study specifically on work-related reading (Adler et al. 1998). The study included fifteen participants from a wide variety of professions. Analysis of the resulting data produced a typology of work related reading.

Of these, four are goal-directed to some degree: 'searching to answer questions' (SAQ), 'reading for cross-referencing' (CR), 'reading to support discussion (DI), and a goal-directed sub-type of 'skimming' (SK) called 'proofing'. Together these account for about 70% of the total reading time (see Figure 14). Hence, goal-directed search is not only important in the search for documents, but also in their use.

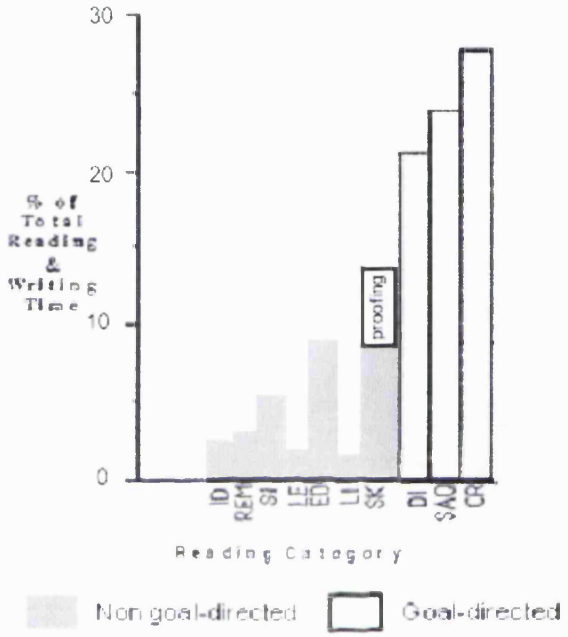


Figure 14 The prevalence of goal-directed, and non goal-directed types of reading, adapted from Adler et al (Adler et al. 1998)

Goal-directed search for relevant spans of text within documents has an important role in many work place tasks such as writing. O'Hara et al conducted an observational and interview study of seven professionals and three doctoral students doing their real-world writing tasks (O'Hara et al. 2002). The study focused on the use of multiple source documents in writing. This uncovered evidence of goal-directed searches such as searching for quotes, integrating material across sources, and referring to ideas and arguments. Similarly, Attfield in an interview study of 25 journalist, reported that journalists print out collections of news clippings to facilitate low cost referencing

during writing (Attfield & Dowell 2003). Thus goal-directed searching and referencing are an important part of the writing process.

Goal-directed search also has an important role in other workplace tasks such as analysis. In the study described at the start of this section O'Day and Jeffries identified six analysis techniques that participants applied to the documents resulting from their searches (O'Day & Jeffries 1993). One of these, "identifying a critical subset" (op. cit., p. 443) of information, clearly comprised within–document goal-directed searching. Russell et al describe a common form of analysis which they term 'Sensemaking' in which goal-directed search plays an important part (Russell et al. 1993). Sensemaking is the process of searching for a representation of data that helps answer task specific questions. They identify this process in a number of different settings from writing research reports to developing a high school algebra course. They identify goal-directed search for information within documents and the mapping of this information to the desired representation as, in most cases, the "most time consuming part of sensemaking" (op. cit., p. 273). Hence, within-document goal-directed search is an important part of many important types of work-place analysis.

The overwhelming emphasis in information retrieval research on between document search seems incongruous next to the prevalence of within-document search. A large proportion of work time is spent on document related tasks. However the vast majority of this time is spent using documents as opposed to searching for them (Sellen & Harper 1997). As argued above, goal directed search within documents is an important form of document use. Hence it follows that people at work also spend a substantial amount of time on goal-directed search within documents. Therefore, research on technologies to support goal-directed search within documents has the potential to make a significant improvement to productivity in the workplace.

4.3.2 The allure of paper

An impassioned call for a paperless office would now seem naïve rather them visionary. Despite contrary predictions (Bolter 1991), paper has far from disappeared; the volumes of paper use in offices is actually increasing (Sellen 1996). Of all the time spent with documents, by the International Monetary Fund workers in Sellen and Harper's study, 56% involved only paper documents and a further 20% involved both paper and electronic documents (Sellen & Harper 1997). Similarly, in Adler et al's diary study 85% of document use was paper based (Adler et al. 1998). It fairly clear that paper is the most prevalent technology for document use; thus it may serve as a good benchmark against which to test the utility of a digital tool for document use. Furthermore, understanding the reasons behind paper's persistent popularity will be instructive in the design of digital alternatives (Sellen & Harper 1997).

A number of studies have attempted to uncover the advantages of paper. One much investigated possibility is that reading from a screen is slower than from paper. Although in the past many experiments have found reading to be slower on-screen, as monitors, software and experimental techniques have improved, the difference is becoming very small (O'Hara & Sellen 1997). Perhaps more important are the abillities of paper to support rapid navigation and maintain context. O'Day et al. observed that for searching "timeliness and brevity" were "referred to again and again by [...] library clients" (O'Day & Jeffries 1993, p. 442); presumably the same is true when single documents are searched. Paper's tangibility and two-handed interaction allow users to search for one page while referring to another, riffle through pages and earmark pages for quick reference. It provides implicit visual, spatial and tactile cues such as thickness for context and rapid, fluid navigation. In a comparison of reading online to reading on paper, O'Hara and Sellen found that on-line users were drawn away from reading by the

"complicated task of dealing with the mechanics" of navigation (O'Hara & Sellen 1997, p. 338). They also found that paper is advantageous because it can act as multiple displays which can be spatially arranged to support rapid access, cross-referencing and persistent organisation of content (O'Hara & Sellen 1997). It can also easily be annotated with highlights and edits, to support collaboration, authoring, and future reference (O'Hara & Sellen 1997; Sellen & Harper 1997). Furthermore, physically it plays an important social role in organisations; documents can be hand-delivered to personalise, aggrandise and assure (Sellen & Harper 1997). Given the multitude of vital affordances offered by paper and the difficulty of reproducing these in a digital alternative, it is clear that paper will be an important workplace tool for the foreseeable future.

However, online document use is also here to stay, so it is important to learn from the advantages of paper to improve online tools for document use. One avenue is to develop hardware solutions for multiple, annotatable, portable digital displays (Schilit, Price & Golovchinsky 1998). Another approach is to develop new ways to enhance the navigability of online documents. The research on the use of paper documents, mentioned above, shows that rapid effective navigation is a valued characteristic of paper. Many within-document navigation tools have attempted to reproduce paper like affordances such as rifling (Woodruff et al. 2000). However perhaps the most effective online navigational affordances will be inherently digital; in this thesis, new digital affordances are explored.

The prevalence and effectiveness of paper suggests that it is an ideal benchmark technology against which to compare new digital alternatives. As argued above, paper is widely, and expertly, used for within-document goal-direct search. Any new technology purporting to support this activity would therefore have to compete directly with paper.

If it is not sufficiently advantageous, it will not be used. Thus when evaluating the potential of a new tool for supporting within-document goal-directed search paper it would be instructive to compare its performance against that of paper.

4.4 Section summary

In this section, existing theories of information seeking have been outlined to argue that formation of information needs and query formulation could be facilitated by a search tool using a metadata taxonomy. It argues that current information retrieval techniques have a number of shortcomings, which for relatively short single documents could be overcome with an approach based on high quality metadata. It goes on to review studies of information seeking and document use in the workplace which are then used to argue that goal-directed search is an important and prevalent form of document use and therefore one meriting digital support. The persistence of paper as a medium for document use is examined to draw out the lessons it has for the design of software support for within-document search.

5 Evaluation methods

Here the different techniques relevant to the evaluation of visualisations that support information seeking, are reviewed. Although information visualisation research has been very active since the late 1980s, until recently there has been relatively little published work on evaluating the resulting technologies (Chen & Czerwinski 2000). This section reviews the relevant work, drawing out the techniques that will apply to the empirical and theoretical evaluations presented in this thesis.

5.1 Empirical techniques

This section concentrates on the empirical evaluation of software for within-document search and the presentation of search results. These evaluations demonstrate the techniques and metrics that can usefully be applied and illustrate some of their strengths and weaknesses. In section 5.1.1, two important evaluations of within-document search tools are discussed. Section 5.1.2 discusses lessons that can be taken from the larger literature on the evaluation of search result visualisations.

5.1.1 Evaluation of within-document search tools

The evaluation of the SuperBook system described above, in section 3.4.1, set a precedent for the evaluation of tools that support goal-directed search within documents. Egan et al report three experiments in which they compared the effectiveness of SuperBook to that of paper for locating answers to questions using a large software manual (Egan et al. 1989). It was thought that the difficulty of answering questions would be dependent on whether words in the questions appeared as words in relevant section titles and paragraphs. Questions were generated from the target information so as to be evenly divided between those with words which matched those in the, title and

text, text only, title only, or neither. After a forty minute training session, ten student participants completed four evaluation sessions. In each one they were asked to search for sections with the answers to each of eight questions; half with SuperBook and half with paper. Once the relevant text was found its section number was recorded. A fifth evaluation session required two essays to be written using the manual, which was presented in SuperBook or in paper form.

The final version of SuperBook had a mean search completion time and proportion of correct responses that was 25% better than with paper. Essays written using SuperBook were also rated higher. The manipulation of question difficulty had the expected effect, when using paper, questions without words in the relevant heading, showed worse performance. Questions without words in the heading or the text, showed worse performance than other question for both paper and SuperBook.

Although in most respects the SuperBook studies laid down the evaluation methods used in subsequent evaluations of within-document search tools (e.g. Crestani & Ntioudis 2001; Hornbæk & Frøkjær 2001; Mynatt et al. 1992; Chapter 5 section 3 of this thesis), some criticisms can be made of this work.

One criticism concerns the peculiarities of materials and questions used. Answers were to be found in only one location. This is not necessarily the case in the majority of documents. Indeed, work for this thesis (see chapter 3 section 3.2.6) suggests that relevant information is often found in multiple non-contiguous locations. When this is the case, accuracy cannot simply be measured by the 'proportion of correct answers', instead the traditional information retrieval precision and recall measures become more appropriate as argued below in section 5.1.2. Another, related, criticism is that in these experiments participants were not subjected to any time pressure. In the workplace, time is often short, so the need for information has to be balanced against the time required to

obtain it. In fact, placing target information at a single location is not ideal for examining the effect of time pressure on retrieval. If target information is at multiple locations a trade-off can be made between the amounts of information found and the time spent. When the target is at a single location this trade-off cannot be made. Finally, each of these experiments demanded up to ten hours of time from each subject. This may be practical when undergraduates are being used as subjects, but would be extremely difficult to achieve in a workplace setting, where potential participants can spare much less time to participate in tool evaluation.

Hornbæk and Frøkjær report an experiment, with a similar methodology, comparing fisheye and overview+detail visualisations to a linear scrolling window (Hornbæk & Frøkjær 2001). This experiment has been described in some detail above 3.4.2, so only the pertinent aspects of it will be discussed here. The goal-directed search task was similar to that in SuperBook in that queries were generated from answers which were to be found in a single location. However, it differed in that written answers were required. These were scored according to the proportion of the points from a model answer that were included in the written answer. As with the SuperBook experiments a large amount of participant time was required (5 hours 15 minutes from each participant). Unfortunately, using this approach yielded no significant differences between interfaces on the goal-directed search task scores, although it did find that the fisheye interface was reliably faster than the linear scrolling window for goal-directed search. The lack of a significant result from the goal-directed search scores, might be due to the increased variability, in comparison to the SuperBook experiment, introduced by requiring written answers to be scored, instead of stating paragraph locations which are either right or wrong.

5.1.2 Evaluation of search result visualisations

The visualisations of document collection search results are attempts to aid users in goal-directed searches. Empirical evaluations of these interfaces offer a number of lessons for improving the evaluation of within-document search tools. The metrics commonly used for evaluating interactive information seeking performance are reviewed in section 5.1.2.1, since they will be used by the evaluations presented in chapters 5 and 6. Section 5.1.2.2 examines a number of less traditional approaches to the measurement of the information seeking performance of visualisation tools and argues that they are not appropriate for use in this thesis. The particular difficulties associated with measuring the performance of information visualisation tools, and the consequences these have for experimental design, are discussed in section 5.1.2.3.

5.1.2.1 Information retrieval metrics in visualisation evaluations

A number of visualisations have been put to the test using the traditional measures of the information retrieval literature namely: recall and precision metrics. Recall and precision are traditionally defined as[3]:

Traditional Recall =

$$\frac{\text{Number of documents deemed relevant by both the system and human judges}}{\text{Total number of documents deemed relevant by human judges}}$$

Traditional Precision =

$$\frac{\text{Number of documents deemed relevant by both the system and human judges}}{\text{Number of documents deemed relevant by the system}}$$

These two measures are important because together they provide a simple assessment of the complementary features of successful retrieval: comprehensiveness, and accuracy

---

[3] In these equations the 'documents deemed relevant to the system' refers to the results set obtained, for a topic, with the information retrieval software; 'documents deemed relevant by human judges' refers to those documents which according to a set of human judges are relevant to the topic.

(Baeza-Yates & Ribeiro-Neto 1999). Although information retrieval systems are generally tested in a non-interactive mode, the major information retrieval evaluation conferences, such as TREC have recently adopted an interactive track (Voorhees & Harman 1998). In the evaluation of visualisations for document retrieval, the human component cannot be separated from the system, and thus slightly different definitions of recall and precision are needed. Veerasamy et al provide an appropriate adaptation (Veerasamy & Heikes 1997):

$$\text{Interactive Recall} = \frac{\text{Number of paragraphs deemed relevant by the user and the judges}}{\text{Total number of paragraphs deemed relevant by the judges}}$$

$$\text{Interactive Precision} = \frac{\text{Number of paragraphs deemed relevant by the user and the judges}}{\text{Total number of paragraphs deemed relevant by the user}}$$

These definitions are often used in the evaluation of search result visualisations, although in print the authors generally fail to explicitly distinguish between the precision and recall measures they are using, and those traditionally used in information retrieval (Ogden, Davis & Rice 1999; Sutcliffe, Ennis & Hu 2000). An additional performance measure that is regularly included in visualisation evaluations is task time.

The widespread use of recall and precision has received much criticism. For large document collections identifying *all* the documents relevant to a query is not feasible (Blair & Maron 1985), thus recall cannot be estimated precisely (Baeza-Yates & Ribeiro-Neto 1999). Another problem is the assumption that there is one relevant set of documents, which should all be retrieved. In reality, different people's relevance judgements have little overlap, and there are often several sets of documents which satisfy the information need (Saracevic 1995). For evaluation of within-document search tools these problems are less serious. Due to the much smaller amount of text involved, recall can be accurately measured. Problems with relevance can be tackled, as

discussed in section 3.1.4 of chapter 5, by using multiple judges and a non-binary relevance scale. Hence, despite the criticisms, precision and recall can be usefully applied to the evaluation of within-document search tools.

5.1.2.2 Alternative approaches to measuring performance

There are many approaches to evaluating interactive retrieval tools which do not involve precision and recall. The evaluations of within-document search tools essentially report the number or proportion of correct answers (Egan et al. 1989; Hornbæk & Frøkjær 2001; Mynatt et al. 1992). These measures however do not capture the trade-off between a search which is accurate (high precision) and one which is comprehensive (high recall).

Morse and Lewis (Morse & Lewis 2000) introduce a sophisticated alternative in which performance is tested on a wide range of visual tasks. Queries are generated from a domain independent taxonomy of visual tasks such as 'distinguish', and 'cluster'. In this way a visualisation is tested on each of the tasks it can support. This is an interesting approach since it has the potential to isolate the precise way in which a visualisation can support search. However, it is not clear whether the tasks generated will be representative of those actually undertaken in the domain being evaluated. Also because Morse and Lewis's (op. cit.) approach requires the testing of a large number of task types, very large numbers of participants are required, which makes it impractical for use in a workplace setting where participants are difficult to acquire.

5.1.2.3 Difficulties of measuring search performance of visualisations

The use of precision, recall and task time in the evaluation of search result visualisation has been plagued by statistically insignificant results. Inter-topic[4] and inter-participant variation can easily overwhelm any differences between the visualisation and the control systems (e.g. Ogden, Davis & Rice 1999; Sutcliffe, Ennis & Hu 2000; Swan & Allan 1998; Veerasamy & Belkin 1996). Veerasamy and Heikes overcame this problem by using a within-participant, within-topic design where relevance judgements were based upon examination of the document title alone (Veerasamy & Heikes 1997).

One interesting outcome of these difficulties is the rationalisations that the creators of visualisations have put forward for the poor performance of their inventions. All the investigators cited above point out the methodological difficulties just discussed. Some attributed the lack of effect to details of the implementation or design; given a better implementation of the idea behind their tools, perhaps better results would have been obtained (Sutcliffe, Ennis & Hu 2000; Veerasamy & Belkin 1996). This exposes a weakness in the empirical approach. An insignificant or negative result from an empirical evaluation is inconclusive; it can be blamed on the visualisation's particular implementation rather then its ultimate potential.

5.2 Theoretical evaluation techniques

A number of concerns have prompted the development of a range of theoretical methods to quantitatively model human performance with software interfaces. Perhaps most influential is the desire to transform HCI into an engineering discipline, with quantitative tools to help shape design decisions. Theoretical tools, can also be used in

---

[4] The description of the information need from which the participant generates their queries.

situations where accessing user populations is expensive or impractical. Subtle effects can be detected and experiments can be repeated with alternative designs and assumptions; all without the need for large numbers of participants. Finally, the internal machinations of theoretical models, which predict or mirror HCI problems can be inspected and understood in detail. Thus they can provide insights into the otherwise inaccessible underlying causes of problems.

In this section the theoretical tools used in this thesis are described and the background from which they arise is discussed.

5.2.1 General principles – rational analysis and cost structure

A general principle that underlies much of the theoretical evaluation tools discussed in the next two sections is that of the rational analysis. This is most clearly stated by Anderson:

> *"General principle of rationality.* The cognitive system operates at all times to optimize the adaptation of the behaviour of the organism" (Anderson 1990, p. 28)

He argues that human cognition has, to some degree, been optimised by human evolution. Thus the key to successfully modelling or explaining cognition lies in determining the optimal adaptation of cognition to the task, environment, and limitations of the human cognitive system. This approach has been successful in explaining a broad range of empirical phenomena in cognitive psychology from memory to choice and reasoning (Anderson 1990). It is an important theoretical underpinning to the theoretical tools described below.

A criticism of the use of this approach in HCI cites the difficulty humans have with software interfaces as a demonstration that they do not in fact behave rationally. However this misses the crucial point that the definition of what is optimal is local to

the user. If learning to use a software interface imposes a large or apparently insurmountable burden, then from the users point of view it might be optimal to use the software crudely or find an alternative means of accomplishing the task.

Another important principle is the notion of cost structure. This concept is particularly useful when discussing information access environments. Generally, in a well-designed information access environment the most frequently required information is the most easily and rapidly obtained, while rarely used information take some effort to find (Card, Pirolli & Mackinlay 1994; Malone 1983); as such the environment imposes a certain structure on the access costs (Card, Mackinlay & Shneiderman 1999). Information visualisations can have their cost structures characterised and quantified in order to pinpoint problems and evaluate alternatives (Card, Robertson & Mackinlay 1991; Pirolli & Rao 1996). Cost structures can be fruitfully investigated using a rational analysis to construct quantitative models such as those described in the next section (Pirolli & Rao 1996).

5.2.2 GOMS a mature engineering modelling tool

GOMS (Goals, Operators, Methods, and Selection rules) is HCI's most widely used quantitative modelling tool (Kieras & John 1996). It was initially proposed developed in the early 80's (Card, Moran & Newell 1983) and since then much work has been done to extend and validate it (Gray, John & Atwood 1993; Kieras & John 1996; Olson & Olson 1990). GOMS models have successfully been used in a variety of situations to evaluate potential interface designs (Kieras & John 1996). Here the principles behind GOMS are outlined, and important differences between types of GOMS technique are discussed.

GOMS is essentially based on the Model Human Processor (MHP) proposed by Card Moran and Newell (Card, Moran & Newell 1983), as a distillation of the psychological

literature into a simple model of human cognition. It essentially proposed nine principles of operation, and three processors; perceptual, cognitive, and motor, specifying times taken for each to function and communicate.

A GOMS model consists of the methods needed to accomplish the specified goals. Each method is a series of steps containing operators. Operators are essentially actions which have not been further decomposed in the model. They change the state of the mental or physical environment and incur a specified time cost. Selection rules are production rules which determine what methods are selected to best accomplish the goal.

There are several variants of GOMS. The simplest, Keystroke Level Model (KLM) (Card, Moran & Newell 1983), does not use selection rules. The analysis specifies a fixed sequence of steps. The KLM provides a single operator, M, for cognitive operators and simple set of heuristics for determining where to place the operators. KLM models have proved to be practical, although not highly accurate, for simple tasks but not suitable for modelling dynamic tasks, since they do not include methods or selection rules (Kieras & John 1996). Card et al also propose a more complete form of GOMS which has been named CMN-GOMS (Card, Moran & Newell 1983). This uses a goal-hierarchy and selection rules, but does not have a clearly defined notation and is thus relatively difficult to put into practice (Kieras & John 1996). Kieras developed and validated a clearly defined alterative form of GOMS named NGOMSL (Kieras 1996). This has been used to accurately estimate performance times, and learning time (Kieras 1996). Finally, a more detailed form of GOMS is CPM-GOMS, was developed to account for the parallelisms that can occur between the processors in the MHP. This is ideal for situations where highly accurate execution time estimates are important and where the users develop extreme expertise; one successful use of CPM-GOMS diagnosed problems with a system for telephone exchange operators who perform

repetitive tasks for which the reduction of the task time by a few seconds, meant savings of millions of dollars (Gray, John & Atwood 1993). Which of these flavours of GOMS is most suitable depends on whether a dynamic model is needed, what level of precision is required and what level of expertise is being modelled.

5.2.3 Other types of cognitive modelling

A number of detailed computational models of human cognition have been implemented and used to model a wide range of HCI and psychology tasks. The three main examples are ACT-R/PM, SOAR, and EPIC (Byrne 2002). These all have their roots in the MHP (Card, Moran & Newell 1983), and are essentially production rule systems. They each have a slightly different focus but are essentially implemented models which embody multiple psychological theories about "those aspects of human cognition that are relatively constant over time and relatively independent of task" (Ritter & Young 2001, p. 3). They have been developed over time to account for a wide range of empirical findings from psychology and can be used to model multi-task performance (EPIC), motor and perceptual performance (EPIC, ACT-R/PM), and learning and problem solving (SOAR, ACT-R/PM). These architectures are sophisticated, and thus difficult, programming environments, which can be usefully applied to evaluations, where fine grain, or complex, human behaviour needs to be simulated. These cognitive architectures are full blown implementations of psychological theory, whereas the GOMS techniques are engineering models, designed to include only the detail required to make predictions about design.

A good example of the application of such cognitive architectures to software evaluation is the work done by Pirolli and Card with ACT-R (Pirolli & Card 1999). As described below, in section 6.3, they adapted ACT-R, adding production rules based on information foraging theory, to model the use of a search tool (Pirolli 1998). In a further

development of their model they have adapted the model to predict eye gaze, and the navigation behaviour of users of the World Wide Web (Pirolli et al. 2002).

5.3 Section summary

This section reviews the empirical techniques used in the evaluation of information visualisation interfaces for information seeking. A review of the evaluations of SuperBook reveals a basic experimental framework for testing visualisations that purports to support goal-directed search within-documents. A brief review the evaluation of search result visualisations identifies metrics appropriate for use with interactive retrieval that have been adapted from traditional information retrieval. The prevalence of statistically insignificant results is highlighted and the implications for the conclusiveness of such empirical evaluations is discussed.

The second part of this section covers theoretical evaluation methods. The incentives for use of theoretical methods are discussed and the main principles behind them are described. The GOMS technique is outlined, and the different forms it can take are compared. Finally, the more sophisticated cognitive architectures, which embody more complete theories of human cognition, are briefly discussed. The next section describes a theory which has been used with one of these cognitive architectures to model information seeking behaviour.

6 Information foraging theory

This section describes a theory, named information foraging theory (IFT), which allows the construction of computational models of information seeking behaviour (Pirolli & Card 1999). Although the traditional modelling tools of HCI, such as GOMS, allow routine actions to be modelled, they are not well suited to modelling the complex decision making involved in choosing between the numerous paths of interaction offered by interactive information seeking tools. Information foraging theory allows concrete predictions to be made about what interaction choices should be made. For this reason IFT has been used in the evaluation phase of the research presented in this thesis. The new evaluation methodology named 'best-case analysis', set out in chapter 6, involves the construction of a computational model to assess the potential performance of GridVis, the tool developed in this thesis to support goal-directed search within documents.

Information foraging theory attempts to explain and predict information seeking behaviour using ideas taken from optimal foraging theory as developed in biology and anthropology. Where optimal foraging theory describes the behaviour of animals looking for food, IFT uses the same principles to describe people looking for information. This section therefore sets out to; describe the aspects of IFT which are used in the research presented in this thesis, present evidence in support of its validity, and describe instances of its previous use in the evaluation of information seeking tools.

The central tenet of optimal foraging theory (on which IFT is based) is that the food seeking strategies of animals are finely tuned to their environment so as to maximise the rate of energy returned per effort invested (Stephens & Krebs 1986). For example, a bird of prey will pursue larger, easy to catch prey (i.e. high energy, low effort) in preference to small, hard to catch ones (i.e. low energy, high effort).

Similarly, at the heart of information foraging theory lies the hypothesis that "when feasible, natural information systems evolve towards stable states that maximize the gains of valuable information per unit cost" (Pirolli & Card 1999, p. 3). This is not to say that people will necessarily develop strategies that achieve a global optimum. Indeed even experts use sub-optimal strategies. This happens through a variety of mechanisms such as the blinkering effect of prior knowledge, and the difficulty of discovering efficient strategies (Bhavnani & John 2001). An optimisation model built according to IF theory describes a possible advantageous strategy that will be exploited if not blocked by other mechanisms such as those above.

More specifically Pirolli and Card assume users are optimal information foragers such that their interaction choices tend to maximise the "valuable knowledge gained from using a system in relation to the cost of interaction with that system" (Pirolli & Card 1998, p.83). From this assumption they elaborate a theory which they claim can be used to make detailed predictions of information seeking behaviour. In order to make these predictions one needs to be able to model how, for a particular interaction choice, users judge the value of information they expect to gain, and the costs of interaction they expect to incur.

The information that enables people to judge the amount of information they stand to gain by following a particular path of interaction is called 'information scent' (Pirolli & Card 1999). Consider a user looking for information on a particular topic in a digital library; they have typed in their search terms and are now inspecting the results page. Their interaction choices are partly determined by an assessment of the amount of valuable knowledge available from, and effort required by, each interaction path before them. This assessment will be based on the information displayed about each option coupled with past experience of using such systems. In this case they must decide which

search results are going to be the most informative and accessible. In information foraging theory, the information that permits this assessment (i.e. the bibliographic data) is referred to as information scent. More precisely information scent is "the (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues, such as bibliographic citations, WWW links, or icons representing the sources" (op. cit., p.10)

Information scent is especially important given that information relevant to a particular need is often distributed in a patchy way. For example, while our digital library searcher might locate some valuable articles and web pages, they will find the majority of the library's contents useless. The valuable parts of the library are referred to as 'patches'; this is a term taken from food foraging (e.g. a rabbit looking for cabbage patches). Foraging can occur both between patches (i.e. finding the relevant parts of the digital library) and within patches (i.e. skimming through a relevant document to find the parts of it that are most useful).

A range of activities can be modelled using IFT, these can broadly be divided into three types: enrichment, scent-following and exploitation. A forager will mould their environment to increase the profitability of their foraging; this is known as *enrichment.* For example our digital library searcher might refine their search terms to produce a better results set. Enrichment will not be covered any further since it does not play a part in within-document goal-directed search.

The forager will use the information scent offered by the environment to seek out the most profitable things to attend to; this is called *scent-following.* The forager will initially use information scent to look for the most profitable patch and then within that patch they will seek out the item of information with the highest scent. Once a suitably

promising information item (say a graph) has been found, it will be *exploited*; the user will process the material and thereby absorb the information required.

6.1 Within and between patch foraging

Unlike food foraging, once a patch of items is chosen these items can themselves be patches. Thus, information foraging behaviour often can be characterised as a repeating cascade of patches and items. This is illustrated in Figure 15; at the top level the user is foraging for an appropriate directory heading in a digital library' contents directory. This is between patch foraging; each directory heading constitutes a patch. At the bottom level, the user has selected a document, and is foraging for relevant sections with the document. This is within patch foraging where the document is a patch, and its sections are items. However, things are not so clear when the user forages for appropriate documents. This could be termed as within-patch foraging, if the directory heading is seen as the patch, or as between patch foraging, if the documents are seen as patches. Thus, whether something is an item or a patch is relative to where the forager is in the cascade.

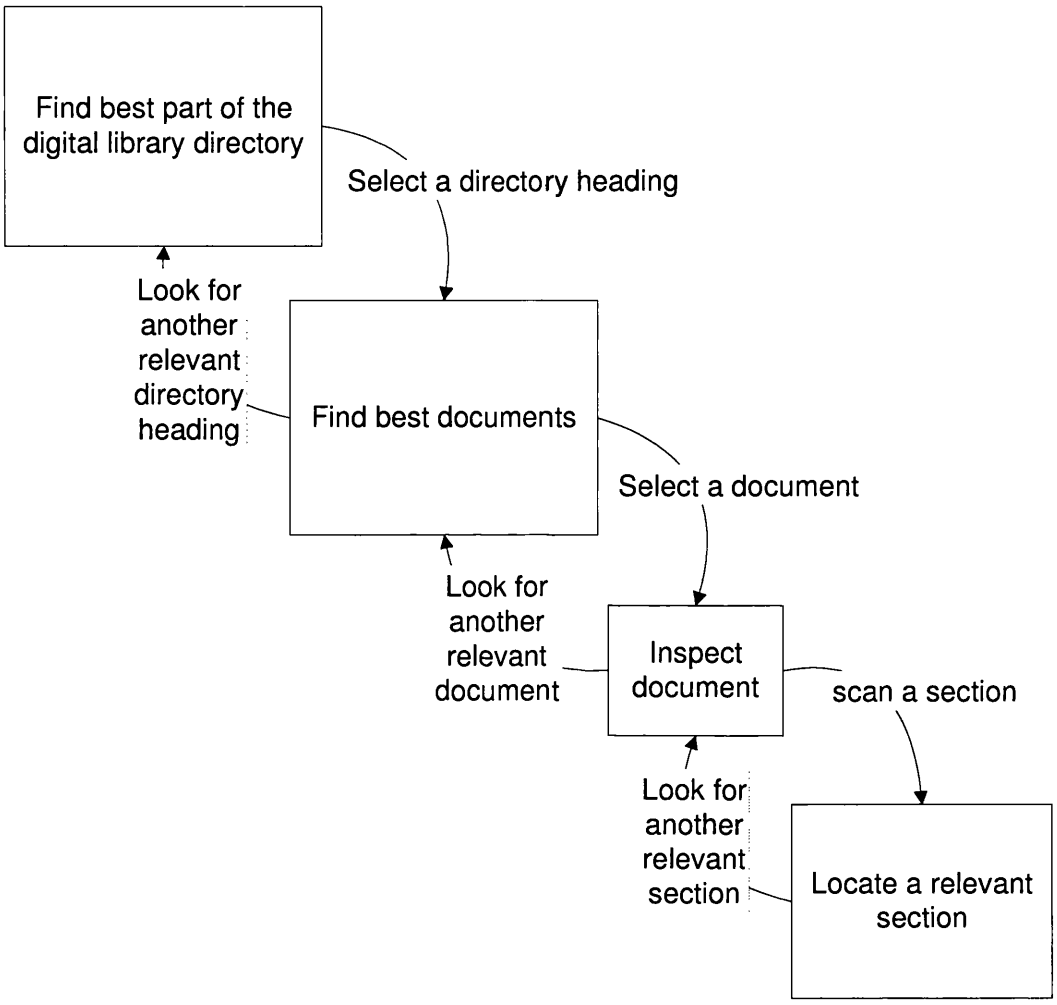


Figure 15 Searching in a digital library; characterised as a repeating cascade of patches and items

An important feature of foraging theory is that it predicts when the forager will transition between levels in these cascades of patches and items. If one does not consider enrichment, foraging behaviour is essentially the same between and within patches. The forager can decide to continue foraging at the current level, i.e. for an appropriate directory heading, consuming items in order of profitability, or jump back up the cascade, and look for a more profitable patch. A jump up the cascade will be made when this increases the overall profitability. This is the case when the next most profitable item is lower then than the mean profitability of the items selected so far. Alternatively, if the profitability of the next patch, i.e. a directory heading, is higher

than that of the next item, i.e. a document, then an optimal forager will abandon the current patch in the next patch's favour.

When within a patch the user must select the most profitable items; the way in which this is done is described by diet models. One such diet model is discussed in the next section.

6.2 The optimal diet selection algorithm

The optimal diet selection algorithm specifies how to construct a diet of information items which, given certain assumptions, maximises profitability (Stephens & Krebs 1986). It assumes that the forager has access to information about the prevalence and profitability of these items and can select them in an all or none manner. It also assumes that the time taken to recognise the items is effectively zero.

In order to explain the rational behind this algorithm it is instructive to start by considering what was discussed above in section 6.1. The user must alternate between spending time, looking for profitable patches, and looking for profitable items in those patches. Given the fact that in information seeking alternative patches, such as items on a results list, are of the same prevalence, the expected time to locate the next patch is assumed to be a constant t_b [5].

Within each patch there will be a number of items. For each item, i, the user can estimate how much information the item is likely to offer, g_i, and how much time this will take to exploit or process, t_{wi} [6]. So for a patch containing k items the total expected

---

[5] Here t_b has subscript b because it is the time taken to do the between-patch foraging. This notation is used in order to maintain consistency with literature on information foraging.

[6] Here t_{wi} is given, w, as a subscript because it is the time taken to do the within-patch foraging. This notation is used in order to maintain consistency with literature on information foraging.

amount of information is $\sum_{i=1}^{k} g_i$, and the total time taken to exploit these items and obtain the information is $t_b + \sum_{i=1}^{k} t_{wi}$. Thus the total rate at which information will be gained from the patch, *R(k)* is[7]:

$$R(k) = \frac{\sum_{i=1}^{k} g_i}{t_b + \sum_{i=1}^{k} t_{wi}}$$

Equation 1 The rate of information gain for a diet of* k *items

If the reader is to maximise their rate of information gain they must select items such that *R(k)* is maximised. To see how this can be done, suppose the user has evaluated the profitability, $_i$, of each item in the patch and ranked them in terms of that profitability. $_i$ is calculated as:

$$\pi_i = \frac{g_i}{t_{wi}}$$

Now suppose the user has selected *k* of the top ranked items, and is considering whether to select the $k+1^{th}$ item. The total information gained from those items would be $\sum_{i=1}^{k} g_i$,

[7] In food foraging theory, t_b varies depending on the type of item the patch contains. This is expressed as prevalence, $_i$, which is $1/t_b$, so that Equation 1 is formulated as $R(k) = \frac{\sum_{i=1}^{k} \lambda_i g_i}{1 + \sum_{i=1}^{k} \lambda_i t_{wi}}$.

and the total time gained would be $t_b + \sum_{i=1}^{k} t_{wi}$ [8]. Then in order to ensure selecting this next item does not decrease the diet's profitability:

$$\frac{\left(\sum_{i=1}^{k} g_i\right)}{\left(t_b + \sum_{i=1}^{k} t_{wi}\right)} \leq \frac{\left(\sum_{i=1}^{k} g_i\right) + g_{k+1}}{\left(t_b + \sum_{i=1}^{k} t_{wi}\right) + t_{wk+1}}$$

This gives:

$$R(k) = \frac{\sum_{i=1}^{k} g_i}{t_b + \sum_{i=1}^{k} t_{wi}} \leq \frac{g_{k+1}}{t_{wk+1}} = \pi_{k+1}$$

Equation 2 The condition that must hold if the* k+1th *item is to be added to a diet of* k *items.

Essentially in order to maximise the profitability of the diet the user must add the items to the diet in order of increasing rank until the profitability of the diet of the top *k* items, *R(k)* is less than the profitability of the next item being considered, π_{k+1}. Thus when the $k+1^{th}$ item is less profitable then that of the diet consumed so far, the user should abandon the item's patch and look for another, more profitable one.

This algorithm provably maximises the profitability of selections within a patch (Stephens & Krebs 1986). It can be summarised as follows:

1. Calculate the profitability of, $_i$, of each item *i*.
2. Rank the items according to their profitability. Where *i* = 1 is the most profitable item and *i* = *n* is the least profitable one.

---

[8] If these items have been exploited then the cost and gain should be calculated using actual values, rather then expected values.

3. Add the items to the diet in order of decreasing profitability until the profitability of a diet of the top k items, $R(k)$, is less than the profitability of the next item being considered, π_{k+1}. In mathematical terms, items should be added to the diet until the condition in Equation 2 is broken.

The way this algorithm works is illustrated by Figure 16. As the items are added to the diet, the profitability of the diet as a whole increases at slowing rate since the profitability of the items decreases. Once including the next item decreases the diets overall profitability, it is not worth continuing. Therefore, the optimal point at which to stop adding items to the diet occurs just before the 'profitability of diet' line crosses the 'item profitability' one.

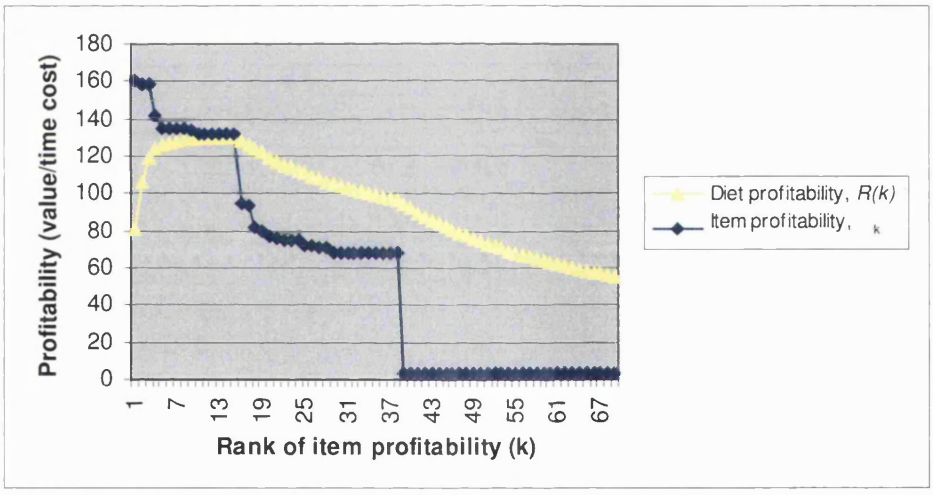


Figure 16 An example of the relationship of between item profitability and the profitability of diets including items 1,2, ... k.

6.3 Constructing models of information foraging

Pirolli and Card have used IFT to build a computation model, called ACT-IF (Pirolli & Card 1999), to simulate the use of a particular information seeking tool. Here we shall very briefly discuss its architecture (see section 5.2.3), and how it models people's

estimation of information scent. These topics are covered here in order to facilitate the discussion later discussion, in section 8.3 and section 2.2.2.1, in chapter 6, about how ACT-IF differs from the computational model of information seeking presented in chapter 6 of this thesis.

ACT-IF instantiates the principles of IFT using Anderson's ACT-R production system architecture (Anderson 1993). As discussed in section 5.2.3 ACT-R is a wide-ranging model of human cognition. IFT can be incorporated into ACT-R as a set of production rules. It is not necessary to discuss all aspects of the ACT-IF model here, however one element, the mechanism for modelling people's information scent estimates will be discussed since it takes a different approach to that used in the research presented in this thesis.

In order to model activation scent judgements Pirolli and Card used a mathematical formulation of the spreading activation theory of memory (Pirolli & Card 1999). The rational for this is that relatedness judgements must rely on the association of items in memory, and this in turn could be the basis for assessments of information scent. Spreading activation theories thus offer a psychologically motivated approach to modelling information scent judgements.

Essentially, the scent of a word is calculated as sum of the Bayesian probabilities of it appearing within 20 words of each of the query words. These probabilities are calculated using co-occurrence statistics from the 2.2 Gigabyte TIPSTER text corpus (Harman 1993). Evidence for the efficacy of this approach is presented in the next section (section 6.4).

However further research found that this psychologically motivated approach to calculating information scent does not show itself superior to a much simpler alternative (Pirolli 1997). In this simpler calculation, scent is based on the sum of the co-occurrence

frequency of each query and word in the document summary. This produces predictions of information scent that are not significantly worse then those calculated using spreading activation. This suggests that despite its complex formulation the spreading activation calculation of information scent is limited by the word frequency data it is based on.

6.4 Empirical support for information foraging theory

Here the evidence that has been put forward in support of IFT will briefly be summarised. Pirolli and Card report the results they obtained from, a testable instantiation of the theory, their ACT-IF model for a particular information seeking tool named Scatter-Gather (Pirolli & Card 1999). Here we shall briefly describe the Scatter/Gather system and go on to look at a few of the instances in which ACT-IF predicted the performance of eight users looking for documents relevant to each of twelve TREQ query topics.

Scatter/Gather is an information retrieval system (Cutting, Karger & Pedersen 1992) that allows a user to browse large document collections via automatically induced document clusters. The user is initially presented with ten document clusters, each displaying the number of documents it contains, and showing a few documents titles as well as a set of keywords summarising the cluster's contents. Users can select those clusters that seem relevant (i.e. *gather)* and then recluster (i.e. *scatter*) the documents in the selected clusters to produce ten new clusters. The user can go on scattering and gathering until one or more clusters seem relevant enough and small enough to merit a more detailed inspection. They can then open these clusters, inspect all the document titles and can select any of these to view the full document.

A piece of evidence that supports the idea that people follow information scent, relied on log files detailing the participants usage of Scatter/Gather. The ACT-IF model is used to make predictions for every user actions. For each screen state ACT-IF produced a ranked list of actions, which were compared to the user's actual actions. One comparison made, was to record ACT-IF's rank of the production which matched the user's choice. The graph below, Figure 17, shows that the productions ranked more highly by ACT-IF more frequently predict actions correctly. Hence, by assuming that users are scent following, IF theory is doing a reasonable job of predicting their choices.

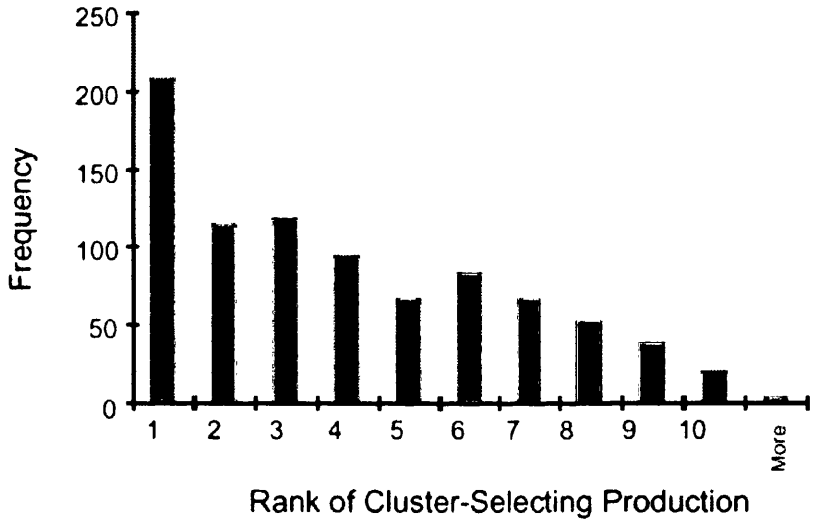


Figure 17 This graph shows that the productions for selecting clusters that were more highly ranked by ACT-IF, were more often selected by users.

Another result which contributes evidence to the validity of the diet model portion of IFT. The optimal diet selection algorithm specifies that a user should select clusters so long as the cluster's profitability is greater than the overall profitability of the clusters gathered up to that point. Pirolli and Card extracted the item and diet profitability estimates from ACT-IF, and calculated the probability the cluster being selected by

users, and that of it not being selected, using experimental log files (Pirolli & Card 1999). Figure 18 plots the results from this. When the line drops below the x-axis, the cluster will not be selected, when above the axis, it will. 0 on the x-axis is where the profitability of the current cluster equals the overall profitability of what has already been selected. So one would expect a high probability of cluster selection just positive of 0, where the item profitability is higher than the profitability of the diet so far, and vice-versa just negative of 0. This is indeed what the graph shows, suggesting that the optimal diet selection algorithm can predict the set of information items users will decide to pursue.

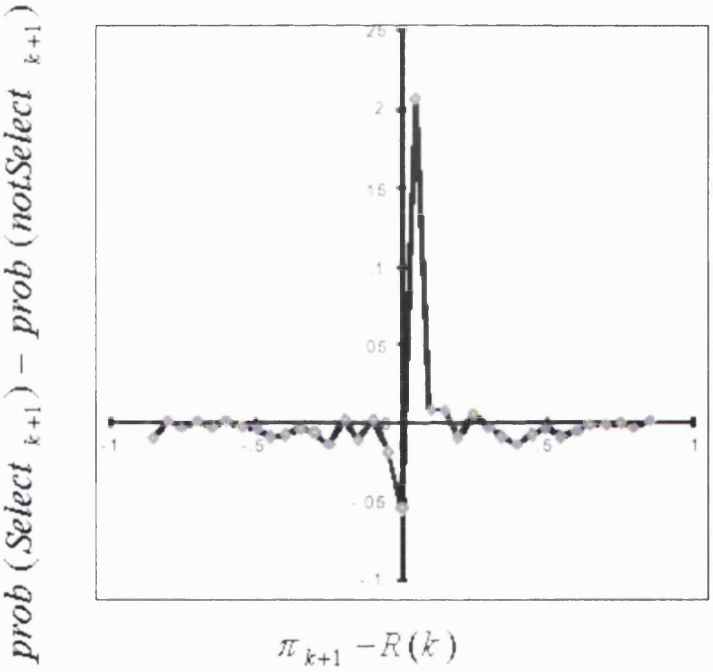


Figure 18 On the y-axis, this graph shows the probability of the next cluster not being selected, subtracted from probability of it being selected. On the x-axis, it shows the profitability of the next item, minus the profitability of the current diet. More simply, whether a cluster will be selected is shown on the y-axis, and whether the profitability for the current item is larger than that for those already selected is shown on the x-axis.

6.5 Section summary

Information foraging theory (IFT) and its use in modelling information seeking behaviour is described. IFT assumes that users will tend towards interaction choices which maximise the value of information gained per unit cost. To do this they use information scent to assess the profitability of alternative information items. The optimal diet selection algorithm, which can be used to model the optimal selection of items, is described in detail. Then information seeking environments are characterised as a repeating cascade of patches (which are collections of items) and items.

The use of IFT to model information seeking in the ACT-IF is described. The use of spreading activation and word co-occurrence frequencies to model information scent is then discussed. Finally evidence for the validity of IFT is put forward drawing on the fit to user behaviour obtained by an ACT-IF model of an information retrieval tool.

7 Assumptions and limitations

This section critically examines the assumptions that underlie the work done in this thesis and their basis in some of the research traditions reviewed in this chapter.

7.1 A narrow focus on information retrieval in general and goal-directed search in particular

This thesis looks specifically at how goal-directed search within documents can be supported. The assumption that underlies, and justifies, this focus is the assertion that goal-directed search within documents is an atomic unit of work which appears in numerous configurations of work. Evidence to support this assumption is described in section 4.3.1, which discusses evidence for the prevalence of within-document goal-directed search in cross-referencing, supporting discussions, skimming, question answering (Adler et al. 1998), writing (Attfield & Dowell 2003; O'Hara et al. 2002) and analysis (O'Day & Jeffries 1993; Russell et al. 1993). However, a further assumption implicit in this thesis is that the requirements for successful support of goal-directed search do not change significantly across these various types of work.

This assumption is somewhat supported by theories of task-based information retrieval and studies of particular work contexts. Far from emphasising the dependence on task type, Vakkari suggests that in task based information retrieval the search tactics and search terms used change as the search process unfolds and the searcher elaborates their understanding of the domain, and of their information needs (Vakkari 2001). By contrast, studies of search in particular work contexts, such as a study of journalistic writing by Attfield and Dowell, and one of lawyers planning jury trials by Kuhlthau and Tama, suggest some aspects of information seeking such as initiation and termination of searches is very strongly determined by the work context (Attfield & Dowell 2003;

Kuhlthau & Tama 2001). However mechanics internal to the search process are much the same as described in general domain independent models of information seeking such as that shown in section 4.1. Thus, existing theoretical and empirical work does not suggest that the structure of goal-directed search changes substantially across work contexts.

Nevertheless these arguments do not deal with the possibility that the requirements for the successful support of goal-directed search can change with the context in general and with work context in particular. Johnson proposes that information seeking contexts can be categorised at three different levels (Johnson 2002). At the 'macro' level changes in society and technology, alter the incentives and opportunities for information seeking. At the 'local' level there are opportunities for and constraints on information seeking; for example in an organisational context their may be a rich technical infrastructure for information seeking, but a strict organisational restriction on the types of information available to individuals. At the individual level differences in technical and domain experience, individual responsibility, time pressure, and motivations can effect search behaviour. Thus, there is much in work context that varies across different types work that can effect search behaviour.

The question that remains is what effect these variations might have on the successful support of goal-directed search within documents. One possibility is that different contexts might lead the user to place differing emphasis on particular qualities such as speed, comprehensiveness, information need clarification, or precision. Thus tools which exceed others in the qualities favoured for a particular context, would have an advantage in those contexts. This argument suggests that the value of a tool in different work contexts can be assessed if the different aspects of a tools performance and the way these trade-off against each other is known. Another possibility is that some

configurations of work involving within-document goal-directed search, might have emergent properties that change the requirements for the successful support of goal-directed search. Attfield and Dowell have recently tentatively suggested that journalists need some way of managing the fruits of their goal-directed searches, i.e. news clippings (Attfield & Dowell 2003). However, even here, the emergent requirement is an addition to the process of goal-directed search, rather than a modification of it.

In summary, although there is much variation across work contexts there is little reason to believe that this changes the structure of goal-directed search. In addition, although this variation in context may determine what constitutes successful support of goal-directed search, it is argued that this largely stems from differing emphasises on particular aspects of performance. Thus, a tool's worth in a particular context can be based on a detailed understanding of its general performance parameters.

A final point is that there are other aspects of information use which could be conveniently supported alongside goal-directed search. For example, information visualisation has been used to support the recognition of outliers and patterns in data (e.g. spotting a predominance in a particular type of information). It would be entirely possible to support this kind of activity with a visualisation such as GridVis, however in the interest of thoroughness, the scope of this thesis is limited to an examination of goal-directed search.

7.2 A focus on expert performance

The final form of evaluation presented in this thesis is a best-case analysis, which is a comparative evaluation of expert performance with paper and GridVis, the tool being assessed. The focus on expert performance rests on the assumption that a tool which

does not lead to superior results with expert use will offer inferior performance across the board. This assumption can be criticised in a number of ways.

Firstly, a new tool may offer superior support for novice users, at the expense of expert users. Although this scenario is commonplace, it does not apply here. Firstly, GridVis is not intended to specifically support novice users. Secondly, there is no pressing need to support novice within-document searchers, since with the alternative technology (i.e. paper) there are no novice users. We can illustrate this point by considering a traditional problem in IR interface research which has been the reluctance of novice users to adopt advanced boolean search techniques (Hearst 1999). Thus, there have been some attempts to produce interfaces, with somewhat restricted capability, which make these techniques accessible to novice users (e.g. Shneiderman & Ahlberg 1994). In within-document search, people almost universally use paper in extremely sophisticated ways. It would not make sense to offer a technology with more restricted capability, that could be made 'easy to use'. Hence, although expert performance may be a misleading basis for evaluation if the tool is aimed uniquely at novice users, the tool developed in this thesis, does not and should not focus on uniquely supporting novice use.

Secondly what does the best-case analysis say about the users that do not achieve the expertise assumed by the model it is based on? The best-case analysis aims to uncover the tool's potential performance. If a tool has a superior potential performance to its rival, it suggests that it is superior at some basic structural level. If this is the case, one can expect that the relative advantage will persist at a range of levels of expertise.

Further criticism might be attached to the way in which expert performance is defined in the evaluation in chapter 6. The best-case evaluation makes use of the rationality principle (see section 5.2.1) and information foraging theory (see section 6) to model

expert performance. This evaluation therefore rests on the assumption embodied in these approaches, namely that human users will tend towards performance that is rational given the bounded nature of their perceptual and information processing capabilities.

One argument against this assumption is that members of an organisation can be expected to align their behaviour with whatever betters their standing in that organisation rather then specifically optimising their performance of work tasks such as goal-directed search. However, normally, organisational incentives are specifically designed to encourage the best performance. A more convincing argument, made in section 6, is that their are a number of factors which can impede the progress of users toward optimal behaviour such as the blinkering effect of prior knowledge and the difficulty of discovering efficient strategies (Bhavnani & John 2001) or more simply the lack of enough opportunities for practice. Nevertheless by ignoring these effects and adopting the rationality principle, the best-case analysis is able to examine the tools potential. If that potential turns out to be significant, then further work can be directed at removing barriers to optimal behavior such as those, listed just above, identified by Bhavnani & John (op. cit.).

7.3 A focus on traditional information retrieval metrics

This thesis principally uses the evaluation metrics typically used in information retrieval (see section 5.1.2.1), namely: task time, recall and precision. These metrics have the advantage of giving results which can be quantitatively compared across systems. A number of criticisms of recall and precision have already been covered in section 5.1.2.1, however an additional one should be mentioned here.

The essential shortcoming of the reliance on these metrics is their failure to touch on some important aspects of how the tool is used and received. They say nothing about other important factors such as, the level of comprehension attained, and the level of enjoyment or satisfaction experienced. However it can be argued that, to some degree, comprehension will follow recall, since a higher recall gives the user the opportunities to absorb more of the relevant information. Also, whereas user satisfaction or enjoyment is an important and generally desirable result of tool use, in a corporate context it is secondary to measures of effectiveness that contribute more directly to the 'bottom-line'.

7.4 A focus on quantitative comparative evaluation

The evaluations in this thesis are essentially controlled comparisons of the goal-directed search performance of alternative tools. They thus look at performance over a single sitting, in a somewhat artificial setting. This limits the kind of information that these evaluations make available.

One argument that can be made is that this type of evaluation only informs us of a tool's effectiveness in a single work context. Since there can be such a large variation between information seeking contexts (Johnson 2002) one can argue that the results of these evaluations are of little significance. In the section above (section 7.1) two consequences of this variation in work context were described.

The first consequence is that different types of performance would be appropriate in different context, placing different amounts of emphasis on properties of the tool such as speed, recall and precision. For example, a business analyst at the start of a big project might value the comprehensiveness, and thus recall, of the search, over speed or

precision. However when rushing to get the facts she needs to make her case in an imminent meeting, speed and precision will be valued over recall. This kind of variation with work context is not problematic for the type of evaluation reported in this thesis, because it permits the properties of the tool to be profiled in detail. This profiling enables examination of the trade-offs between important performance parameters such as precision, recall and speed.

The second consequence of variation in work context is that there may be unanticipated but important requirements that emerge when goal-directed search is embedded in particular types of work. In order to study these emergent requirements and their implications other types of evaluation are necessary.

Another approach to evaluating the tool would be to introduce it to a variety of work contexts and observe its use. This type of study would be longitudinal, in that it would allow the observation of usage over weeks and months. This would permit an examination of the question, discussed above in section 7.2, of the degree to which users become expert users. It could also be a contextual study that gathers data on how the tool is fits into the wider work context (Beyer & Holtzblatt 1995). This would allow problems of integration with other tasks to be identified, and the emergent requirements to be identified. Unfortunately this approach has a number of downsides. Firstly, it does not produce data that allows the direct comparison of the tool to its competitors. Secondly this type of study is extremely difficult to initiate, since it requires the researcher to wield a substantial amount of power in the organization, in order to introduce the use of a new tool before its benefits have been demonstrated.

8 Comparative review of related work

This section synthesises research on information visualisation, hypertext and the evaluation of interactive systems, comparing and contrasting the systems, techniques, and methodologies developed in these areas to those used for this thesis. The design work of this thesis builds on techniques pioneered in a number of different information visualisation systems; these systems and the techniques they develop are discussed in section 8.1. Research into hypertext and e-books essentially aims to support online reading; as such it has resulted in techniques to support within-document search. In section 8.2, these techniques compared to those developed for this thesis. Section 8.3 reviews a number of evaluation techniques and contrasts them with best-case analyses, a new evaluation technique put forward in chapter 6 of this thesis.

8.1 Visualisation

A number of visualisation systems have attempted to support navigation through and search within documents. There has also been some work on the visualisation of metadata taxonomies. The majority of both these types of visualisation have not been subjected to any evaluation, nevertheless it is possible to distil out and critically assess the principle techniques used. Section 8.1.1 covers visualisations which present a miniaturised image of the document as a navigational affordance. Section 8.1.2 looks at tools which visualise the distribution of content across the document. Section 8.1.3 examines visualisation techniques used to present the organisation of metadata. In each of these sections, the techniques developed in previous work are critically evaluated with reference to the requirements of visualising paragraph-level document metadata.

8.1.1 Navigation with visual document overviews

Much work has looked at how an *entire* document can be rendered on screen. Eick's SeeSoft System (Eick, Steffen & Sumner 1992) simply reduces each line in the document to a line one pixel high in a narrow vertical rectangle. An alternative approach was taken in the document lens (described in section 3.4.2) which allows the user to rapidly scan through a document. It uses a distortion techniques to allow the text of interest to be read while keeping the entire document in view (Mackinlay & Robertson 1993). Holmquist points out that the use of distortion makes the text outside the focus area harder to recognise or read, and imposes a large computational burden (Holmquist 1997). He proposes an alternative in which the pages outside the focus area are represented as thumbnails arranged before and after the page currently displayed at full size. The user then linearly riffles through the document using GUI buttons. Ginsburg et al present a related technique in which thumbnails of each page are fitted into a vertical tray by overlapping them (Ginsburg, Marks & Shieber 1997). Users can uncover thumbnails by passing their mouse over them, and view the text by selecting a thumbnail.

Although there has been no direct evaluation of providing navigation and overview via the miniaturisation of a document's visual features, subjective assessment and the prevalence of this technique suggests that it is effective. Also the work of Holmquist and Ginsburg et al, outlined just above, shows how this technique can be extended to deal with long documents. Miniaturisation of a document's visual features is thus a technique which is integrated into the document visualisation tool developed for this thesis.

8.1.2 Navigation and search using visualizations of content distribution

Another innovation made by Eick was to indicate the distribution of content using colour highlighting of elements within a visual document overview (Eick, Steffen & Sumner 1992). This technique has been used in several subsequent within-document visualisations. In the Document Lens visualisation, which is described more fully in section 3.4.2, sentences containing a searched for string are highlighted in red (Mackinlay & Robertson 1993). The PopOut Prism, described to section 3.4.1 (Suh et al. 2002) allows the user to enter numerous search terms which are each highlighted in both an overview and the main document using colours chosen to be maximally distinctive from the colours in the document. In the Reader's Helper, also described in section 3.4.1, Graham highlights words that are relevant to topics of interest to the reader (Graham 1999). Here there can be many possible topics, and the highlighting styles are left up to the user. This can result in highlighting styles that are difficult to distinguish from each other.

As described in section 3.4.1, a user study of the PopOut Prism interface found that the display of keyword distribution and density provided useful affordances for document search. This suggests that a visualisation of paragraph-level metadata should also display such information. However a drawback with these techniques is that they use colour to indicate keyword location. This means that only a limited number of keywords can be visualised, since if there are many colours it is difficult to distinguish between them (Ware 2000) or hold them in working memory. These techniques could not therefore be applied to the visualisation of paragraph-level metadata where there are many keywords; documents prepared for this thesis can have upwards of seventy pieces of metadata each.

The TileBars visualisation, described in section 3.4.1, could overcome this problem, and allow a large number of document metadata items to be visualised. Documents are divided up into segments which discuss a topic and the relevance of a query to each segment is represented by a square; the darker the square the higher the relevance of that segment. The squares are arranged in a row to represent the entire document. Several separate term sets can be incorporated at once by visualizing the relevance of each with a separate row of squares. Whereas colour alone restricts the number of metadata items that can be easily identified, perhaps a visualisation based on TileBars might be used to show the distribution of a large number of items. For this reason, the visualisation of paragraph-level metadata developed for this thesis, extends the TileBars technique, adapting it to be capable of showing the distribution of a large number of metadata items.

8.1.3 Visualizations of metadata taxonomies

There has been work on making it easy to navigate through and understand metadata taxonomies. The vast majority of this work has focused on visualising the taxonomy's tree structure. The paragraph-level metadata developed in this thesis is housed in a taxonomy. The aim of displaying the taxonomy is to aid the user in locating metadata of interest. If the taxonomy is to fulfil this aim, an effective visualisation of its structure is required. Thus, techniques that have been used to visualise the hierarchical structure of metadata are briefly reviewed with the aim of identifying the technique most suitable for use with visualisations of paragraph-level metadata.

A number of information visualisation systems, described in section 3.3.2, have been developed to visualise the very large subject categories taxonomies used in some document collections. Risden et al visualised a catergory hierarchy with over 10,000 nodes using XML3D, a three dimensional hyperbolic browser (Risden et al. 2000).

Hearst and Karadi used the cone tree to visualise, and Schneiderman used a 2-D horizontal tree to visualised the even larger MeSH category hierarchy. Finaly, Benjamins et al. use a hyperbolic tree to visualise an ontology for metadata used to annotate web pages (Benjamins et al. 1999).

The techniques used in these visualisations are focused on how to handle large hierarchies. By contrast, it must be noted that the taxonomies needed for paragraph-level document metadata are fairly small. When the visualisations used for these large taxonomies have been subjected to evaluations (see section 3.1.2.1), some have proved effective, e.g. XML3D (Risden et al. 2000), and others, e.g. the Cone Tree (Cockburn & McKenzie 2000) and the hyperbolic tree (Van Der Wege & Pirolli 2000), have not. However, when the taxonomies being visualised are not extremely large, the widely used vertical intended folding tree performs as well as, or better then, other tree structure visualisations (see section 3.1.2.1).

8.2 Hypertext and e-books

Hypertext covers a very diverse field of research. Here comment is restricted to techniques that have been developed for searching or navigating through single documents. One aspect of this is the attempt to develop e-books that are a real alternative to their time tested paper ancestors. This has resulted in much research into enhancing legibility, developing hardware platforms (Schilit, Price & Golovchinsky 1998) and automatic conversion to electronic format (Landoni, Crestani & Melucci 2000). However what is of interest here is the work both from the e-book and hypertext literatures on software to support navigation and search within documents. This research is reviewed in order to highlight that this thesis investigates those areas in which are most in need of further research. One successful branch of work, covered in section 8.2.1, has extended the traditional table of contents functionality in various ways, to

create a range of outlining technologies. Another branch of work discussed in section 8.2.2, has explored less structured ways of linking together related information.

8.2.1 Outlining

Outlining is a powerful technique for supporting within-document navigation that has been integrated into widely used software products such as help systems and word processors. It was invented as part of the SuperBook system, and its effectiveness has been amply demonstrated by a number of experimental evaluations (Egan et al. 1989). As such it is not a technique on which further research is needed. Outlining is described, the evidence for its effectiveness is presented and the potential advantages, for goal-directed search, of paragraph-level metadata over outlining are discussed.

The SuperBook system, described in detail in section 3.4.1, displayed a detailed table of contents to left of the document window. This could be expanded to reveal varying levels of heading at points of interest. This view was integrated with the highlighting of search words similar to those discussed in section 8.1.2 above. A set of empirical evaluations, found SuperBook to be significantly faster than paper, and it is highly likely that the outlining functionally played an important part in this success.

More recently, outlining has been used to present long web pages on personal digital assistant (PDA) screens (Buyukkokten, Garcia-Molina & Paepcke 2000). The hierarchical structure of a web page is automatically inferred, and phrase or keyword summaries are generated for each point in the hierarchy. When a document is viewed, summaries for the top level of the hierarchy are displayed. These can be successively expanded until the full text is revealed. A comparison of this system to the standard web browser available on the Palm™ personal digital assistant, found it offered substantial advantages for the narrowly defined goal-directed searches examined. Brown and Weihl

report a similar, although less advanced, application for desktop browsers (Brown & Weihl 1996). A more sophisticated desktop system was reported by Boguraev et al (Boguraev et al. 1998). It combines the thumbnail style document visualisations, described in section 8.1.1, with temporal typography (Wong 1995), to present summaries, generated by summarisation techniques, of selected points in the document

The use of outlining has been shown to be effective for desktop and PDA screen. It has also been widely adopted commercially and a number of avenues for extending the technique have been explored. In a corporate context, outlining technique has some disadvantages. Outlining essentially follows a documents content and structure. Therefore, the view it presents cannot be tuned to support invariant information needs. In a corporate context there are information needs which are relatively constant across the organisation and within its sub-groups (Bryce 1997). Metadata, on the other hand, can be designed to fit the needs of a user group; content of interest can then be viewed through the metadata, concentrating the user's attention on the information and patterns of interest. Thus while outlining has undoubtedly proven its effectiveness, paragraph-level metadata may offer advantages over outlining for the work-place user.

8.2.2 Unstructured hypertext navigation techniques

The Dexter hypertext reference model serves as comprehensive description of the capabilities aspired to by many existing hypertext systems (Halasz & Schwartz 1994). In it, hypertext is defined as a network of nodes, which for our purpose correspond to chunks of text. Interestingly, the model expressly avoids defining the structure or navigation mechanisms that apply within nodes. This is indicative of the emphasis in hypertext research on the navigation between nodes. As such, in general, hypertexts are effectively multiple document collections. However some research has been done on rendering single documents as hypertexts. These essentially make use of a number of

navigation techniques. One capitalises on the hierarchical internal structure of documents to produce the outlining based navigation aids discussed in section 8.2.1. Two more traditional hypertext navigation techniques are, links between related nodes and topic indices where each topic is a one-to-many link, i.e. a link to multiple nodes or anchors within nodes. Finally, many hypertexts have integrated full text search facilities such as those discussed in section 3.4.1. Here the traditional navigation techniques offered by hypertexts and their application to within-document search, are discussed.

Marchionini and Shneiderman report a study comparing performance on goal directed searches using a hypertext system, named Hyperties, with that using a paper version of a fifty two page maintenance manual (Marchionini & Shneiderman 1988). Hyperties offered, an index, links embedded in the text, and a 'back' button. This study found that performance was significantly slower with hypertext but that accuracy was not significantly different. A later study repeated this experiment with a 411-page encyclopaedia and a more fully featured hypertext system which added a full text search tool and two way links (Mynatt et al. 1992). Overall hypertext was slightly more accurate but also slightly slower; hypertext is only substantially superior for factual questions where query words do not match titles of relevant sections. More recently a system was evaluated which offered an index, dynamically generated for any term selected, on top of the features possessed by the hypertext systems of Mynatt, Leventhal, et al. and Marchionini and Shneiderman (Crestani & Ntioudis 2001). Crestani and Ntioudis found that multiple-choice questions were answered more accurately but less rapidly with hypertext, while for open-ended questions paper was more accurate but less rapid. Hence hypertext systems have only been shown to exceed performance with paper in some respects. Whether paper or hypertext do better on

accuracy or speed depends in complex ways on the type of question and the nature of target material.

The systems above offer the user a number of navigation tools and it not clear which the users tend to use. However Marchionini and Schneiderman report studies showing that indexes are faster for goal-directed search than links embedded in text (Marchionini & Shneiderman 1988). These systems also offer text searching, dealt with in section 4.2, outlining dealt within in section 8.2.1, and the one-to-many links in indexes. One-to-many links function in much the same way as metadata; they are terms which link to several points in the text. The research outlined above on hypertext systems with one-to-many links suggests that these can be effective for navigation, but their lack of consistent superiority over paper suggests that improvements would be desirable. The research in this thesis looks at new ways of using one-to-many links for within-document search. Specifically the visualisation technique developed shows the distribution of anchors locations and co-occurrence of metadata at those locations.

8.3 Theoretical techniques for performance evaluation

This thesis proposes the best-case analysis, a new theoretical approach to evaluating the potential of a new application. This involves using the GOMS and information foraging theory modelling techniques, to compute the best performance that could be achieved by a human user of the tool under evaluation. This performance is then compared to the best-case performance, derived empirically or theoretically, of a competing alternative. This section reviews a number of the best-case analysis' predecessors. These have emerged from Xerox PARC which has pioneered the use of theoretical analysis to determine performance of interactive software since the publication of Card's seminal book 'The psychology of human-computer interaction' (Card, Moran & Newell 1983). The various types of theoretical performance evaluation used to evaluate software are

reviewed and their shortcomings in relation to the kind of evaluation performed for thesis are discussed.

One theoretical approach to comparing the performance of systems is to compare their cost structures by determining their 'cost of knowledge functions'. This defines how the number of items accessible varies as a function of cost in time (Card, Pirolli & Mackinlay 1994). This approach was used to compare two calendar programs. Empirical measurement and regression analysis were used to estimate the time taken to access a day as a function of how distant it was from the current date (Card, Pirolli & Mackinlay 1994). Since the visualisation being evaluated was slower than the competition for all but very distant dates, design improvements were considered. Parameters in the cost of knowledge function were changed to simulate the effects of proposed enhancements; this revealed that the visualisation would still under perform even with the improvements. The essential limitation of this approach is that in many information retrieval applications it is the informativeness of retrieved items which matters more than the sheer number that are readily available. As such it does not offer a particularly informative approach to examining the cost structure for satisfying an information need.

GOMS has been used to compare the performance of an application to existing competition and explore the performance of alternative designs. The support of two packages (Splus and TableLens) for exploratory data analysis tasks was compared by predicting the time taken for a range of elementary tasks (Pirolli & Rao 1996) such as finding a variable's median and judging the shape of its distribution. GOMS models were constructed to predict the time taken to do these tasks with each package as a function of the size of the variable set being examined in the task. These models of elementary tasks were then combined to give a model of two higher level tasks, (e.g.

'finding the important features of the variables'). These were used to plot graphs of the change in task time for each of the packages, as the task is repeated with an increasingly large set of variables. This comparison demonstrated that the information visualisation tool, which the researchers had developed, offered faster or comparable performance to the existing Splus package.

In general, the GOMS models used in these comparisons represent best-case performances: the best methods are selected for accomplishing each task, performance is error free, and the time estimates are derived from measurement of well-practiced routine tasks. Crucially the tasks modelled with GOMS are entirely straightforward in that the best procedure can be specified. This is unlike many information retrieval tasks involving information visualisations. In these, a set of interactions leading to the goal form one of many possible paths through a large interaction (or problem) space. A visualisation's ability to inform the user about which paths are shortest and most fruitful is key to its success. This ability must be tested by any theoretical tools that purport to evaluate visualisations that support information seeking.

One approach to dealing with the large interaction spaces inherent in information seeking tools was reported by Pirolli (Pirolli 1998). This used information foraging theory and dynamic programming to determine the optimal path through the interaction space. This approach was used to explore two hypothetical improvements to the Scatter/Gather system, described in 6.4, under different task conditions. These improvements and task conditions were incorporated into models of the Scatter/Gather system that defined the costs (time) and benefits (number of relevant documents) of the various interaction choices. Dynamic programming was used to search through the interaction space for the most profitable interaction path. By this means, a direct comparison between performances attained for each proposed improvement can be

made. Any performance differences observed, result directly from differences in the design or task conditions. For example, Pirolli found that improving the clustering quality by 25% produced superior performance when compared with doubling the clustering speed. However, when there was only a short time in which to complete the task, doubling the clustering speed produced better performance. This demonstrates how the coupling of dynamic programming and information foraging theory can be used to calculate an absolute optimum performance achievable with an interface. However, it takes no account of some important limitations of the interface and of the human operator. Any information seeking tool displays limited amount of information scent for used in making interaction choices. Thus, unlike a dynamic programming algorithm, users are limited in the accuracy of their estimates of information value and the likely costs associated with alternative options. The actual interaction paths chosen will depend on the nature of the information scent displayed. Since an approach based on dynamic programming ignores this factor, and calculates the *absolute optimum performance*, it is not suitable for computing the upper bound on the performance a human operator of the tool might achieve. For this reason a dynamic programming approach cannot be used in a best-case analysis, since in a best-case analysis the theoretically calculated performance, is compared to *empirically measured expert performance* on a competing tool. For a best-case analysis a theoretical approach is needed which can calculate the upper bound on performance achievable by a human user.

There is, however, an existing technique, which could be used to compute the upper bound on performance achievable by a human user of an information retrieval device. This is typified by the integration of information foraging theory into ACT-R to create ACT-IF (see section 5.2.3 for more on ACT-R). This has been used to model the usage

of scatter-gather (Pirolli & Card 1999) and produces a reasonable fit to user data (see section 5.2.3). It was therefore decided to use the same approach in the best-case analysis described in this thesis. Here, however, it was not necessary to use ACT-R, since its comprehensive scope is not needed. The model will only be to applied to one tool and thus only some aspects of the user need to be modelled. In contrast, Pirolli and his colleagues at Xerox PARC, need ACT-R's comprehensive scope since they intend to model a wide range of information seeking applications and phenomena. For example they have recently been extended ACT-IF to model eye gaze in web page use (Pirolli et al. 2002).

8.4 Section summary

A number of visualisations are reviewed which attempt to aid navigation with miniaturised images of the document such as those used in GridVis, the visualisation tool developed in this research. Although there is no direct evaluation of the utility of such images, its wide adoption suggests that is effective. Techniques for visualising content distribution are reviewed. It is concluded that the keyword highlighting technique is not able to clearly represent a large number of different keywords at once, as is required for paragraph-level metadata. The TileBars visualisation can however be extended for use with an arbitrary number of items as is done in GridVis. The successful use of outlining to support search and navigation in a number of within-document navigation tools is described. The disadvantages of outlining relative to paragraph-level metadata for work-place users are highlighted. The navigational affordances developed in research on hypertext are discussed. The similarity of the one-to-many links, used in the indexes of hypertext books, to the use of paragraph-level metadata for business documents, is noted. The mixed results obtained with this form of link suggests that research on alternative representations, such as the visualisation of paragraph-level

metadata in GridVis, might be desirable. Thus, the techniques developed for within document search and their relationship to the work done in this thesis have been reviewed.

The theoretical evaluation techniques used in the construction of expert performance are reviewed. Approaches used to compute a tool's cost structure are problematic in their concentration on number of items obtainable per unit cost. This excludes any notion of item quality and cannot be used to calculate performance metrics for comparison with empirical data as required for a best-case analysis. The use of GOMS is reviewed, but it is noted that alone it is not able to simulate the selection of paths through the large interaction spaces involved in information retrieval devices. The use of information foraging theory would allow this to be accomplished. However, when used with dynamic programming the absolute optimum performance is calculated rather than the upper bound on the performance of a human operator. A cognitive modelling approach, as typified by the use of ACT-R, can be used to simulate human 'best-case' performance. It is argued however that the use of the full ACT-R architecture is not necessary for the simulation of performance on a single tool since the model will not need to be adapted for use with other tools. Thus in the best-case analysis described in chapter 6, information foraging theory is used with GOMS models to model human best-case performance.

The theoretical evaluation techniques used in the construction of expert performance are reviewed. Approaches used to compute a tool's cost structure are problematic in their concentration on number of items obtainable per unit cost. This excludes any notion of item quality and cannot be used to calculated performance metrics for comparison with empirical data as required for a best-case analysis. The use of GOMS is reviewed, but it

is noted that alone it is not able to simulate the selection of paths through the large interaction spaces involved in information retrieval devices. The use of information foraging theory would allow this to be accomplished. However, when used with dynamic programming the absolute optimum performance is calculated rather than the upper bound on the performance of a human operator. A cognitive modelling approach, as typified by the use of ACT-R, can be used to simulate human 'best-case' performance. It is argued however that the use of the full ACT-R architecture is not necessary for the simulation of performance on a single tool since the model will not be adapted for use with other tools. Thus in the best-case analysis described in chapter 6, information foraging theory used with GOMS models to model human best-case performance.

Chapter 3 The production and use of corporate documents and how metadata could fit in

1 Introduction

The previous chapter combined a diverse range of background literature to lay the foundation for the research in this thesis. The two studies reported in this chapter were conceived to extend this foundation.

This chapter describes two studies that examined the production and use of documents and the possibilities for the production and use of metadata. These studies were carried out at the head office of Sainsbury's, which is a large UK retailer. The studies provided an understanding of the needs of document users and the constraints of their environment. They also helped examine two aspects of the feasibility of metadata use and production. Specifically, can the information needs of readers be represented and supported using metadata, and can a document's writer bear sole responsibility for metadata production?

The type of metadata of interest here, is the 'subject heading' metadata consisting of keywords or noun phrases. Subject headings are a typical form of semantic document metadata (Chapter 2 section 2.2). Research concentrated on this type of metadata since it might be an effective basis for content visualisations (chapter 2 section 3.3.1 and 3.4.1). Although grammatically rich text might also be used for visualisation, the automated extraction of semantic relations from syntax is a research topic in its own right, and is thus considered to be beyond the scope of this thesis.

In the first study, the 'writer study', a range of Sainsbury's employees were interviewed about their role as writers of business documents. In the second, the 'reader study', those who read the documents involved in the writer study were interviewed about their roles as readers of documents. These studies are exploratory; their primary use is to guide further work. For this reason, the studies were designed to gather a broad range of data and the sample sizes are relatively small.

2 The writer study

This study was intended to gather a number of different pieces of data. These fall into two distinct camps: the context of corporate document use and the feasibility of using metadata at the within-document level. First, in order to study the context of document use, data was gathered on what documents writers produced, and how they went about writing them. Second, to examine the feasibility of using metadata, the writers were asked to list the goals they had in writing the documents and those they thought their readers had for reading them. This data, in conjunction with that gathered from the reader study (see section 3), was used to examine the accuracy of the writers' understanding of their readers information needs. An attempt was made to reformulate the goals identified using metadata, in order to assess the feasibility of creating metadata that aligns with reader goals.

2.1 Method

The data was collected using semi-structured interviews. Seven participants were chosen to cover a broad range of documents dealt with in the Logistics Division of Sainsbury's. As such, the participants represented a range of functions within the business. One was a statistician, two worked on knowledge management, one on change management, one on IT systems development, and the remaining two worked on systems and supply chain management; three of the seven were managers. An effort was made to keep the male to female ratio roughly even.

The interviews were conducted using a variety of interview techniques. First, participants were asked to focus on differences between cases[9]. This, along with the use

[9] This is a technique used in repertory grid knowledge elicitation (Gaines & Shaw 1993)

of a range of the participants' documents, encouraged them to talk about concrete examples rather than vague generalities. Second, a suggestion from Holtzblatt and Beyer was used; participants were asked to elaborate on the more abstract things they said in order to get them to talk about concrete examples (Beyer & Holtzblatt 1995). Finally, the general parameters of and approach to, the interview was based on the 'long interview' as used in anthropology (McCracken 1988). This specifies the kinds of questions needed, and how the interviewer should behave, in order to put the interviewee at ease and obtain the data desired.

The interviews were recorded onto audiotape, and notes were taken by the interviewer during the interviews to keep track of what had been covered. The documents brought along by the writer were numbered and photocopied; each participant was asked to refer to the document by its number during the interview so that it could be identified from the audio recording.

When the interviews were arranged each of the participants was asked to bring five to six documents they had written to the interview. At the start of the interview the participants were asked if they could go through each of the documents that they had brought along, focusing on the type of documents they were, and the differences in terms of what they were trying to get across to their intended audience. Next they were asked to describe how long it had taken to write each of the documents, and how they had gone about doing so. The participants were encouraged to draw out the differences between their writing processes with each of the documents. Once they had completed this, the participants were asked to go through each document and describe its official and unofficial goals.

Each participant was then asked to pick a single document for which the interviewer could elicit: the participant's goals, the people who read it, and what the writer expected

them to need from the document. The participants were asked to choose a document still reasonably fresh in their memory which they would be prepared to discuss in detail. They were then asked to list those who had read the document. For each reader, the participant was asked to describe the interests or goals that they would have in reading the document. Here the study purposely did not focus on *information* needs, but inquired about the participants' *goals* for a document. This was done because corporate documents do not exist solely to satisfy information needs; political, social, and process issues could also be important. They were then asked to summarise these goals using key words. Finally, participants were asked to point out which paragraphs in the document each reader would actually need to read.

2.1.1 Methods used for the data analysis

The first step was to go through the audiotapes after each interview and take detailed notes whilst referring to the copies of the documents being spoken about. Different questions required slightly different forms of analysis. However, a dominant approach was to examine the detailed notes and produce a simple categorisation scheme relating to the questions or hypotheses of interest and relevant literature. The scheme was then used to mark up the text in the detailed notes, and the occurrences of each category were counted up. These counts were used to test hypotheses or identify patterns and point to parts of the data in need of further examination and categorisation.

2.2 Results and discussion

This sections starts by describing the methods used to analyse the data collected. What the data had to say about each of the various questions tackled in this study is described with a separate section being given to each major question. Namely:

1. What kind of documents are produced?

2. How are they written?
3. Can the goals people have for consulting documents be represented by metadata?
4. Does this metadata pull together information dispersed in non-contiguous regions of the document or does it simply replicate existing document organisation?

Due to the broad scope and exploratory nature of this study, some of the results obtained pertain to writing practices; these are omitted here as they are not relevant to this thesis. Those that remain are included in section 2.2.1.

2.2.1 The document types produced at the Sainsbury's logistics planning division

The data gathered was used to find out what kind of documents were generally produced and used. Table 1 shows that the majority of documents were reports and short memo documents. The documents can be roughly divided into non-prose, and prose document types. The prose documents are more prevalent, although they too, often include substantial non-prose components, tables of numbers and graphs. They are almost all written in terse short bullet points, each making fairly discrete points. This style of writing is different from an academic or literary writing style in which points are always threaded along an argument or narrative. Another important feature is that the documents are often richly structured, and tend to have at least two or three section headings per page.

		Numbers of documents of different lengths (num pages)										
		All lengths	1	2	3	4	5	6	7	8	9	10 +
Prose documents	Report/memo	24	5	6	3	2	1	2	1	1		3
	Project plan	2						1				1
Non-Prose documents	Tabular information resource	4	2			1		1				
	Task list	2	2									
	Process diagram	4	2	1		1						
		Total frequency	11	7	3	4	1	4	1	1	0	4

Table 1 The prevalence of different document types and the frequency of different document lengths

2.2.2 Can goals be expressed as metadata?

The goals that the writers thought their reader's had were examined to assess whether these might be convertible to usable metadata. More specifically, a goal was considered convertible to usable metadata if it could be expressed using less than four key words. Such goals will be referred to as *metadata expressible*.

There were two properties that contributed to making a goal not metadata expressible. Some goals were too *evaluative*. These required the readers' evaluation as to what information was relevant, and are thus beyond the power of writer produced metadata (e.g. "What would the system implications be of the current proposal?"). Other goals were too *diffuse*. These were only satisfied by information that was diffusely woven into the fabric of the document (e.g. "What is the final form of the ideas that have been discussed?"). A simple scoring scheme, shown in Table 2 and Table 3, was devised to quantify the prevalence of these attributes.

Some evaluative goals might be broken down into sub-goals which are metadata expressible (e.g. "What would be the system implications be of the current proposal?" -> "Does the proposal involve the XYZ system?"). Nevertheless, the extent to which

metadata can represent goals without them having to be broken down first, will go some way towards determining its usefulness.

Evaluativeness level score	Description of level of Evaluativeness
1	All or almost all of the material relevant to the goal can be specified by the writer
2	The bulk of the material relevant to the goal can be specified by the writer
3	Only some of the material relevant to the goal can be specified by the writer
4	None or almost none of the material relevant to the goal can be specified by the writer

Table 2 Scoring scheme for evaluativeness

Diffuseness level Score	Description of diffuseness level
1	None of the material relevant to the goal is scattered throughout the document
2	Some of the material relevant to the goal is scattered throughout the document
3	All of the material relevant to the goal is scattered throughout the document

Table 3 Scoring scheme for diffuseness

The schemes in Table 2 and Table 3 were applied to all the goals the writers predicted their reader might have. The results of this can be seen in Figure 19. A goal is taken to not be problematically evaluative, if it has an evaluativeness score of 1 or 2. Relevant text must also be isolateable, hence a diffuseness level score of 1 or 2 is necessary. Given these criteria, 64% of the goals were metadata expressible.

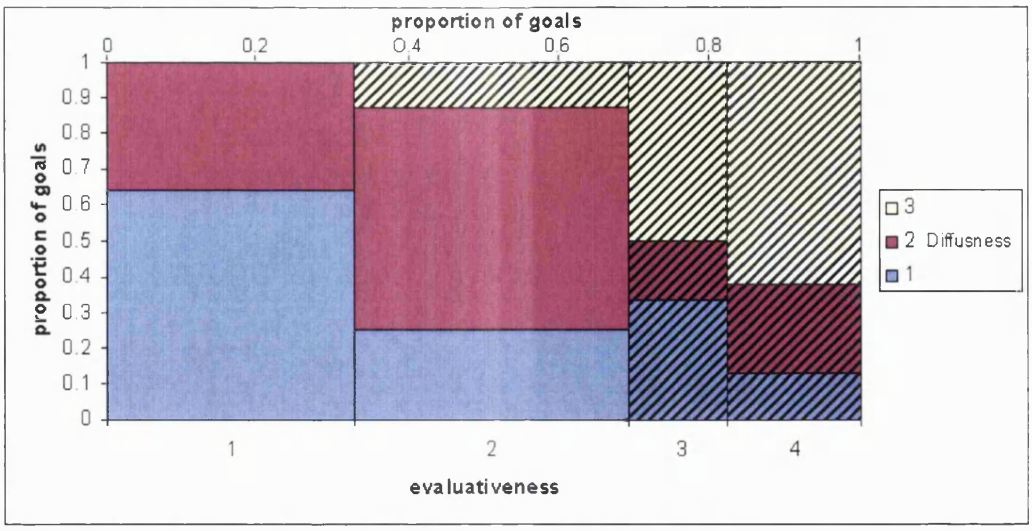


Figure 19 The proportion of goals with each level of diffuseness and evaluativeness rating. The portions of the graph that are hashed out indicate those goals which are not metadata expressible.

2.2.3 Is the metadata replicating existing document organisation?

Another feature of the goals important to the feasibility and usefulness of metadata is the degree to which the relevant pieces of text are scattered across the document. If pieces of text relevant to a metadata term are not contiguous then the metadata performs the useful function of relating information which the document structure does not co-locate. Conversely, if the information related to metadata is all located in one contiguous chunk, then the usefulness of the metadata is partially undermined. To look at this issue the experimenter counted the number of contiguous spans of text relevant to each goal proposed by the writer for their readers. Relevant text was found to be in one contiguous span of text for only forty percent of the goals. Hence, for sixty percent of the goals the relevant text was in multiple contiguous blocks. These goals had largely the same distribution of evaluativeness and diffuseness values as those reported in section 2.2.2 above. This suggests that metadata could indeed be used to bring together informative parts of a document to satisfy information needs.

2.3 Summary

In the writer study seven Sainsbury's head office workers were interviewed about the documents they had recently authored. Data was gathered about what types of documents they wrote, and how much time they spent doing so. Documents were on the whole, short and richly structured taking less than a day to write. Data was also gathered about the goals authors had for their documents and those they thought the documents readers would have. Analysis of these goals established that over sixty percent of them could be expressed as metadata. Further analyses showed that this metadata would not simply replicate existing document structure since the majority of it related to multiple non-contiguous spans of text.

3 The reader study

In this study the main goal was to gather data which would assess the predictions made by writers from the previous study about the information needs of their reader. To this end four people were interviewed who had read one of the documents covered in depth by one of the writers from the previous study. They were asked what their goals had been in reading the document, and which parts of the document had been useful to them. These were then compared to the predictions made by the document writer.

These interviews were also used to identify and examine scenarios in which metadata might be used. Thus, the interviews also looked at: the range of documents the participants read and reused, how they went about getting the information they needed and what got in their way when doing this. The interviews concentrated on those documents that were used repeatedly over time and contained highly valued information. This was done because, if metadata improves access to information, it is for these documents that the investment in the creation of metadata will bring the highest payoffs.

3.1 Method

As in the writer study the data was collected using semi-structured interviews. The choice of participant was based on the data collected in the writing study. They were selected from the readers of those documents that had been looked at in depth in the writers study.

Only four participants could be included in this study. This was due to problems of availability and the criteria that the participants needed to meet. The participants needed to be recipients of information rich documents with long useful lives whose writers had given a full set of goal predictions in the previous study. The participants selected had

actually read these documents as part of their work. Although the selection criteria imposed constraints on the size of the reader study it ensured that genuine behaviour could be uncovered rather then idealised generalisations (Beyer & Holtzblatt 1995). The participants were split between genders, management and non-management, and the Logistics and Information Systems divisions.

	Logistics	Information systems
Management	M, F	M
Non-management	F	

Table 4 The characteristics participants used; M signifies a male participant, and F signifies a female one.

Each interview started with the participant being presented with the document for which their goals had been predicted in the last study. They were asked to describe each of the information needs they brought to the document and to summarise these in three or four word bullet points. In the previous study, the participants had been asked to summarise goals using key words. This did not have the effect of eliciting usable metadata. Asking for three or four word bullet points turned out to produced a much better result.

Next, each participant was asked to go through the document with a pencil marking out the paragraphs they had read and those containing information they needed. They were also asked about their opinion of the document in order to gauge the document's success.

The next portion of the interview was intended to examine the types of documents and scenarios in which metadata might be used. It made use of the ten to fifteen recently read documents which the participant had been asked to bring to the interview. The first step was to isolate the documents most rich in information with the longest useful lives. Each participant was asked to go through the documents and select those, which they or someone else, would want to refer back to in the future. Then the participant was asked

to look through each of these documents, explaining what kind of document it was and marking out the sections that they would most likely reuse. In order to gather information about reading praticies, the participants were then asked to go through each document and discuss how they went about extracting the information needed and the difficulties they encountered in doing so.

3.2 Results and discussion

This study, like the last one, produced results looking at a number of questions, each of which will be treated in separate sections. Namely:

1. What kind of documents are read?
2. How are they read?
3. How are they reused?
4. Do document authors understand their readers' needs?
5. Can the goals people have for documents be represented by metadata?
6. Does this metadata pull together information dispersed in non-contiguous regions of the document or does is simply replicate existing document organisation?

3.2.1 The characteristics of long life, information rich documents

The creation of metadata will generally be most worthwhile for documents that have long useful lives and are information rich. Hence, by considering what kinds of documents these are, document types suitable for enrichment with metadata will be more easily identified.

The participants were asked to consider the documents they had brought along and select the five with the longest useful life and most information rich. These documents

were looked at and compared to each other in order to look for common characteristics. Table 5 shows the result of this distillation of the common characteristics of long life, information rich documents.

Document characterisation	No of documents (Total=19)
Specifications	8
Fact records	5
Table-based document (e.g. projects register)	3
One off reports	3
Cyclical documents (regularly reused and updated)	3
Dynamic documents (e.g. task lists)	2

Table 5 The prevalence of different long life, and information rich document types

The most common document type is specification documents. This kind of document is highly standardised and structured. They primarily serve as a record of decisions of what needs to be done or built. The prevalence of this document type is partially due to two of the four participants being from the Information Systems Department. For many systems, this kind of specification document is the only substantial piece of documentation and is therefore an important repository of corporate knowledge. Although these are already highly structured they are often very long and might be enhanced with within-document metadata.

The next largest set of documents are primarily repositories of data and information. Some of these are dynamically updated from databases. Another important group of documents are regularly used and updated management tools such as project registers and task lists. These are generally in the form of short or highly structured lists or tables. Within-document metadata is therefore unlikely to enhance them since the information they contain is already highly accessible.

3.2.2 Reading practices

The participants were asked to go through each of the documents they had selected as long lived and information rich, and describe how they had gone about reading it and what problems they had encountered. Table 6 is a characterisation of the reading styles of each of the four participants.

	Participant 1	Participant 2	Participant 3	Participant 4
Knows the structure of documents and uses this to determine reading strategy and what sections to read	Yes	Yes	Yes	Yes
Annotates		Yes	Yes	
Skims	Yes Uses the visual texture of the page content to locate interesting material (e.g. graphs) Process: 1. Decides whether a document is relevant by looking at title, then perhaps management summary or purpose 2. Goes through entire document reading headings. 3. Before reading a paragraph will look at title to decide whether to skip over it	Yes Process: 1. Reads through it all, starting with management summary 2. Goes to section of document that covers the targets they are concerned with and look though first sections to make sure that they understand what is being done.	Yes Different reported strategies are: • Reads the whole document, skimming the parts not relevant to her team ("read with lower levels of concentration") • Reads the first bit, then skims through looking for relevant sections	Yes Process: 1. Skims the entire document the first time because cannot predict exactly what sections will be needed 2. Goes back to concentrate on important sections
Other		Many of the documents are used in meetings		

Table 6 Characterisation of the reading styles for each of the four participants

There were some important features that were shared across the participants. First they all made use of their knowledge about typical document structures to determine which sections to focus on. Three of the four also used skim-reading techniques to locate

relevant information. Two of them used section headings while the other skimmed through the entire text. A final important feature is the presence of different reading strategies which they have developed for different document genres, and for different tasks.

One important question centres on whether these documents were being read in a goal-directed manner. The readers reported that they were attempting to satisfy a number of goals when reading the documents. So it is likely that the skim reading tactics described in Table 6 were an attempt to sample those portions of the text that related to the reader's goals; hence the reading could be viewed as goal-directed. This agrees with the findings of Adler et al. who found that this kind of reading made up a large proportion of workplace reading tasks (Adler et al. 1998).

3.2.3 Characterising document reuse

The participants were asked whether they would reuse documents and if so how. It was found that the way documents were reused seemed to depend on document genre. These patterns of reuse are outlined in Table 7. The table shows the document type, describes the pattern of reuse and whether the re-use is goal-directed[10].

[10] Goal-directed reading is defined as "sampling information in the text which satisfies the goal of the search" (Adler et al. 1998)

Document type	Reuse pattern	Is the use goal-directed?
Specification documents	When something goes wrong with a system and the implementers are no longer available these documents are used to figure out what the problem might be.	Yes & No (can be read linearly for overall understanding)
	Resolve contractual disputes (i.e. was X meant to be done or not).	Yes
	During the implementation these documents are used to look up production goals.	Yes
Process documentation	Used when the process is performed; followed the instructions step by step.	No
Research documents	Used in preparation for large tasks, - presentations, organising an event. Read the whole document for inspiration and background.	No (read linearly for background knowledge)
Fact records	Used in a very modular fashion. i.e. *look up* particular facts.	Yes

Table 7 Characteristic document reuse by document type

Although Table 7 outlines the examples of reuse the participants spoke about, what was surprising was the *lack* of reuse. It seems that most of the documents are read once and rarely looked at again. It is difficult to know whether this is because they are genuinely not needed. It is certainly true that someone who does need to read a document, but is unaware of its existence, would be unlikely to be able to find it with the current level of Sainsbury document (and knowledge) management. Looking for documents relevant to a business problem is not a practical strategy. Nevertheless, much of the reuse that is reported would seem to involve goal-directed search. This is important since it is precisely this kind of search which could be supported by a metadata based navigation tool (chapter 2 section 4.1).

3.2.4 Writer's understanding of the reader information needs

The writers' predictions of the needs readers' claimed to have were compared to their readers' information needs. This comparison enables the traditional information retrieval (IR) performance metrics of precision and recall to be calculated[11].

This study looked at two different aspects of the writers understanding of their readers' needs were studied to assess the possibility of writers being responsible for metadata production. How well do they understand their readers' goals, and thus what content their documents need? Second, given the content they have written, do writers know which portions of it will interest a particular reader? If writers have a good understanding of the goals their readers' have for a document, then they might usefully express them as metadata. If the writers know what parts of the document their readers' need, then they could use metadata to mark them out, thus leading the reader straight to the relevant parts of the document.

The writers performed rather badly overall when predicting the goals the reader would have. The mean precision was 0.31 and the mean recall was 0.38; the variance is high (see Figure 20).

[11] The results have been expressed using the terminology of IR namely precision and recall which have been fully defined in section 5.1.2.1 of chapter 2. However, in brief, recall is calculated as the proportion of chunks which the reader claimed to need that are correctly predicted by writer. Precision is calculated as proportion of predicted chunks that were needed by the reader.

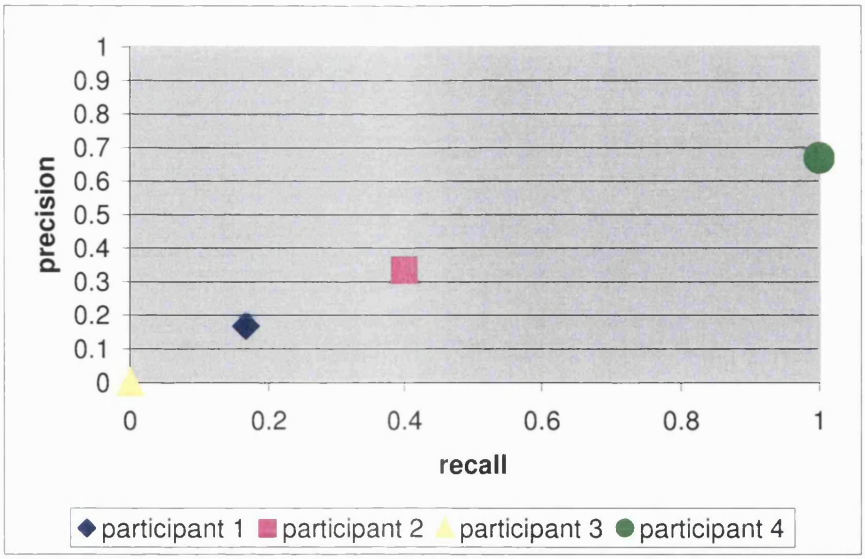


Figure 20 Recall and precision, for the goals predictions made by writers of four documents (one reader/participant for each document)

To make sense of the results it is necessary to note the effect of the very small number of goals offered by two of the participants. Participant 4 offered 2 goals, and participant 3 offered only 1; the other two offered 6 each. The small number of goals for participants 3 and 4 makes their precision and recall scores very volatile. Nevertheless, the low recall and precision evident in Figure 20 suggests that overall writers are not very accurate in their predictions of the goals of particular readers.

It is possible that the recall would increase, if the entire set of predictions made by the writer (rather then those specifically targeted at an individual reader) were used to predict the reader's goals. An analysis showed that this resulted in increased recall (0.42 instead of 0.38) with an accompanying decrease in precision (0.25, instead of 0.31). The increase in recall is not large enough to raise it clear of the high variability, let alone to satisfy a potential reader of these documents. Therefore, in conclusion, it seems that writers do not make very accurate predictions of their readers' goals.

These results conform to findings from the literature on the psychology of writing. In a synthesis of this literature Hayes reports an experiment which found that the clarity of instruction manuals could be significantly improved if the writers were given think-aloud protocols of people trying to use what they had written (Hayes 1996). He cites another study which finds that these kinds of benefits seem to generalise to subsequent writing tasks. Hayes' explanation for these results is that writers use themselves to model the needs of their audiences.

The next question is, can writers be trusted to point out parts of the document relevant to particular readers? The writers managed to predict the portions of text needed by each reader with a mean precision of 0.5 and recall of 0.6. Although the variability is again high (see Figure 21), the performance is generally better than that achieved for goals. Nevertheless it is highly unlikely that readers of business documents would be satisfied by being directed to only 60% of the material they want 50% of the time. Hence, the results suggest that writers cannot make satisfactory predictions of what parts of the document their readers will need.

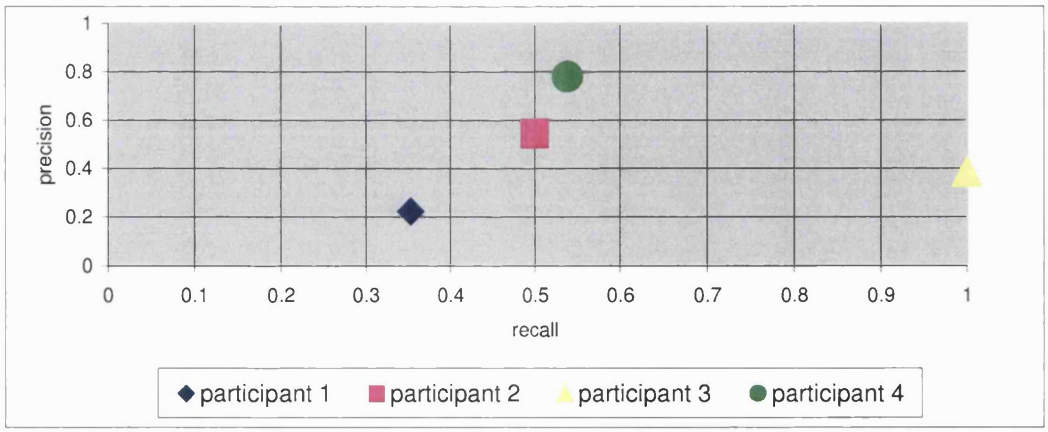


Figure 21 Recall and precision for predictions of the portions of text needed by reader

The findings on the ability of authors to predict what portions of text will be needed by readers is also consistent with findings from the literature on the psychology of writing. Hayes et al tested the hypothesis that authors use themselves to model the needs of their audiences (op. cit.). An experimental group were given some information relating to a text, and a control group were not. Both groups were asked to underline the parts of the text that they thought would be unclear to other people. The experimental group were less likely to identify parts for which they had received supplementary clarification than the control group. This lends weight the hypothesis that people use themselves to model audience needs and illustrates how this might lead to inaccurate predictions of the portions of text individual members of the audience might need.

The findings of the reader study coupled with evidence from existing literature suggest that writers of business documents cannot reliably predict the information needs of their audience. This would seem to preclude the possibility of writers producing paragraph-level metadata solely on the basis of consideration of their readers' needs.

3.2.5 Can goals be expressed as metadata?

The degree to which reader goals were metadata expressible was examined using a similar analysis to that done for the writer study. The results, shown in Figure 22, suggest that 80% of the reader goals are metadata expressible. This result taken together with that from the writing study suggest a worthwhile proportion of a reader's information needs might be encoded using metadata.

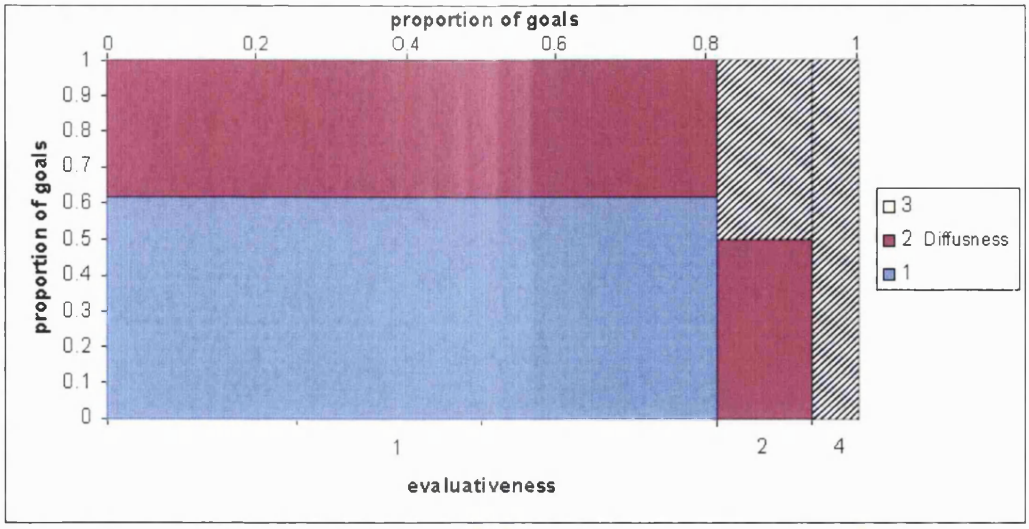


Figure 22 The proportion of goals with each level of diffuseness and evaluativeness rating. The hashed portions of the graph indicate those goals which are not metadata expressible.

3.2.6 Is the metadata replicating existing document organisation?

As in the writer study, the reader goals and the information that was relevant to them were examined to see if the metadata would allow associations of information that were not already grouped together by the document structure. The result was very similar to that obtained in the writer study; for 68% of goals, relevant material was to be found in multiple non-contiguous spans of text. Also, as in the writer study, these goal had a similar distribution of diffuseness scores, although the evaluativeness scores were lower; 80% of goals are metadata expressible. This adds further weight to the conclusion that metadata could bring together dispersed spans of text that relate to a particular information need.

3.3 Conclusion

These studies have touched on many aspects of document production and use. This conclusion highlights those aspects that were most pertinent to this thesis.

The studies provided a basis for the characterisation of documents suitable for enrichment with metadata. These are relatively long (above 4 pages) and take a reasonable amount of time to write (3-4 hours). They are information rich and have a long useful life. In addition, they are written in short paragraphs and bullet points and are generally factual rather then argumentative in style. These documents are particularly suitable for enrichment with within-document metadata, since they consist of fairly small spans of text, that are self-contained units of information, which can usefully be tagged with metadata. In addition, the text relevant to satisfying, over half of the reader goals, was scattered across non-contiguous portions of the document. This suggests that metadata could be useful in bringing together the information relevant to particular reader goals. This characterisation of the Sainsbury's document landscape suggests that there is an identifiable portion of documents to which within-document metadata might usefully be applied.

The characterisation of reading strategies and document reuse indicates that much document use is goal-directed. Since in chapter 2, section 4.1, it is argued that paragraph-level metadata could support goal-directed search, this adds weight to the supposition that paragraph-level metadata might significantly benefit corporations in general, and Sainsbury's in particular.

The feasibility of using metadata was further confirmed by an analysis which looked at whether the goals of document users could be expressed using metadata. The analysis suggests that at least half the goals expressed by the reader and the writers could be represented using key words or noun phrases as metadata. Consequently, metadata might be used to support a useful percentage of information needs.

An analysis of the writers' understanding of the readers' information needs, suggests that writers should not be solely responsible for the writing of metadata. Writers were

found to make poor predictions of both the goals their readers had and the portions of the documents they would need. Hence, it is unlikely that writers could themselves produce metadata which speaks to their readers' needs and points to relevant text. An alternative arrangement would be to characterise and incorporate the readers' needs in a representation upon which the writers could then base their production of metadata. One possible representation is a taxonomy which captures readers' needs as terms or short phrases, and organises them hierarchically by topic.

To summarise, the two studies described in this chapter suggest that paragraph-level document metadata might be a suitable basis for an extra navigational affordance to supports goal-directed search in certain types of business documents.

As a result a tool was designed and built to support goal-directed search via the interactive visualisation of paragraph-level metadata. The design and construction of this tool, and the metadata it uses, is described in the next chapter.

Chapter 4 GridVis: a tool for supporting goal directed use through within-document metadata visualisation

1 Introduction

The previous chapter put forward evidence to suggest that metadata could be the basis for a tool to support goal-directed search within some types of business documents. Here the development of just such a tool is described. This tool, named GridVis, uses paragraph-level metadata to generate a visualisation of a document's content with the aim of supporting goal-directed search and navigation.

To assess the merit of using metadata in supporting goal-directed search, paragraph-level metadata and accompanying taxonomies were prepared for two sets of documents. For the first set of documents these were handcrafted using established methods. A number of improvements and adaptations to these methods were devised to create the metadata for the second set of documents.

GridVis was developed in several stages, moving from concept sketches and paper prototypes, to several live prototypes. Formative evaluation including formal and informal user testing, controlled experimentation, and qualitative and quantitative analysis, was used to warrant design decisions. These evaluations are described in the next chapter, but their value in progressing the GridVis design are documented here.

The development of GridVis and the metadata taxonomies it visualises relied on informal user testing with a particular group of potential users. These were the members of the team supporting this project. This team were Sainsbury's head office workers.

They were an articulate and enthusiastic group of collaborators who were genuinely part of the target user group, conscious of its motivations, knowledge and limits. For the sake of conciseness, this group will be referred to as the 'frontier users'.

In this chapter, the methods used to construct the metadata and taxonomies are laid out, GridVis is described, and various key design decisions are discussed using evidence from analyses, user trials, and experimental evaluation.

2 Creating Content-Descriptive Paragraph-Level Metadata for Business Documents

In order to properly create and evaluate a visualization of paragraph-level metadata, one needs to have high-quality metadata. If the metadata is not of high quality then any negative evaluation of the visualization might be attributed to the shortcomings of the metadata. If the visualization is found to be valuable, then its efficacy with more scalable (and steadily improving) methods of metadata production can be assessed.

Methods for the creation of metadata lie on a continuum; at one end of which low-quality metadata can be produced in a scalable manner with relatively little human involvement, at the other end high-quality metadata is painstakingly hand crafted. Here the methods used to produce the high-quality hand crafted metadata needed for GridVis are described.

2.1 Selecting the documents

In the course of this research, GridVis was trialed with two distinct types of internal Sainsbury's documents. First analytical research reports were used, then later the system was trialed with weekly industry news digests.

The analytical research reports were used partly because they were documents with appropriate characteristics and partly due to practical concerns. These documents conformed fairly well to the characterisation, described in Chapter 3, of documents suitable for enrichment with paragraph level metadata; they were of reasonable length, took more than three hours to produce and were information rich documents and with long useful lives. What is more, they were documents produced by the frontier users, which meant that excellent access to the documents and their authors was afforded. Since this was the first set of documents to be enriched with metadata, the authors

played an important role in validating and checking the quality of the metadata produced.

The second set of documents were chosen based on the experience of working with the analytical research reports. As discussed in Chapter 5, section 3.4, the importance and complexity of the narrative in the analytical research reports reduced metadata's effectiveness. For the second set of documents, the chosen type of document avoided this problem. The weekly industry news-digests documents, are made up of news clippings; they have little narrative structure. However, except for their short authoring time, they do conform to the characterisation, set out in chapter 3 section 3.2.1, of documents suitable for enrichment with metadata. Although the relatively short authoring time might make producing the metadata an onerous additional investment, this might be offset by the increased return on this investment due to the large audiences for this document type.

2.2 Metadata Design and Production

The metadata was produced in what can broadly be broken down into two stages: an initial generation and a subsequent paring down or rationalization. This was done on two separate occasions for the two different document types. For the first set of documents, both stages were executed using methods drawn from the ontology literature (i.e. Uschold & Gruninger 1996; Jokela, Turpeinen & Sulonen 2000). For the second set of documents, these methods were enhanced in order to produce metadata more suited to supporting goal-directed search.

2.2.1 Tag Generation

In creation of the metadata for the first set of documents, a number of techniques were adopted from the literature. In the first stage of the process used, the documents'

audience was examined, as suggested in Jokela et al., to gather information about requirements (Jokela, Turpeinen & Sulonen 2000). The authors were interviewed to discover who the audience were, what their jobs entailed and how their interests differed. The next stage was akin to Uschold and Gruninger's brainstorming stage (Uschold & Gruninger 1996). Here two of the documents were read and liberally tagged with the aim of generating a broad initial set of tags. This initial set was then complemented by reading domain-relevant documents such as business plans and systems documentation, and by attending key team and division meetings. The combined use of these methods led to a large and rich initial set of metadata tags. However, although this tag set represented the important concepts and objects of the domain, they were not necessarily those needed to discriminate topics of interest.

The purpose of the metadata is, of course, to support goal-directed search. To do this it has to enable the user to *discriminate* relevant from irrelevant content. The major shortcoming of the methods used for the first set of documents is that they produced tags which weren't necessarily those needed to discriminate content of interest. The tag set may cover the important concepts and objects of the domain, but these may not be best suited to the task of discriminating relevant material in documents. To deal with this shortcoming, a number of methodological improvements were made, to the tag-generation process use. As above, domain knowledge was required, but since the research continued in the same business environment, the tagging process could rely on the domain knowledge already gained from the investigations of the previous set of documents. The improvements to the method for producing the metadata were primarily derived from a set of eight interviews with a reasonably broad variety of typical readers. These will be referred to below as the 'information needs interviews'. Participants were asked what drew them to the documents and what they looked for when reading them. Each interviewee was also asked to go through one of the documents deciding what they

would skip, which paragraphs they would read and what questions they would be trying to answer. This part of the interview provided a rich source of information about the document audience's needs from audience members themselves, instead of the document authors used in the production of metadata for the first set of documents.

These information needs interviews were also used to generate tags that were tightly focused on supporting goal-directed search. Since goal-directed search involves finding text relevant to a goal, it should be possible to judge the relevance of a piece of text using tags alone. The information needs interviews were used to tease out the kinds of information used in making relevance judgements when reading the types of document for which the tags were being developed. Interviewees were asked to articulate the rationales behind their choices of paragraph, i.e. what features of paragraphs (or sections) prompted them to read in depth or move on. These rationales revealed the minimal set of information, and hence metadata, needed to successfully discriminate material of interest. Moreover, the rationales, and hence the metadata they provide, were expressed in the everyday language of potential users. For example, one paragraph was read because "I always want to know what Tesco is up to, because they are our direct competition", so from this we see that the tag 'Tesco' could be fruitful, and more generally, that the competition and especially the direct competition need to be tagged. Once tags had been derived from these rationales, the insights gained were used to generate more tags through a liberal tagging of all the documents. Hence data from interviews with document audience members allowed the generation of kinds of tags needed to make relevance judgements.

2.2.2 Tag applicability

Each tag applied to a document was given the additional attribute, namely, applicability. In this section this additional attribute is explained and warranted.

Not all paragraphs will involve a topic (say, 'distribution') to the same degree. One paragraph may mention in it in passing, while in another it may be the central focus. Hence, in a sense, the tag 'distribution' is less applicable to the former paragraph, and more to the latter.

This variation in applicability is likely to be particularly important for within-document metadata and goal-directed search. Paragraphs, to which a tag of interest is more highly applicable, are likely to have more to offer. Applicability is also useful when interpreting the meaning of a combination of tags. For example, Table 8 shows the metadata for two paragraphs. If a user were to search for Tesco's financial results, paragraph 2 would be a more promising choice.

Tag	Applicability for paragraph 1	Applicability for paragraph 2
Financial results	High	High
Tesco	Low	High
Asda	High	Low

Table 8 Tags and their applicability to two paragraphs

Each tag was assigned an applicability of between nought and three (see Table 9). This is similar to the approach taken in the PICS standard, in which a number is assigned to indicate to what degree a descriptive term applies to a web resource (PICS 1996).

The decision to use the four-point rating scale, shown in Table 9, was largely driven by pragmatic considerations. The purpose of an applicability rating is to indicate the different degrees to which a tag describes a paragraph. Also, since applicability is not a highly differentiated concept, it is unlikely that a broader scale would deliver any more information.

Applicability level	Description
3	Is principally about
2	Is partly about
1	Mentions in passing
0	Not related

Table 9 The meaning of each applicability level

2.2.3 Tag Rationalization

Once an initial set of tags had been generated, using the techniques described in section 2.2.1, the next stage, for both sets of documents, was to refine and reduce the tag set to a functional minimum.

Following Uschold (Uschold 1996), the tags were inspected for duplication and consistency. The tags were also inspected and reworded for clarity by both the researcher and frontier users.

One other feature used to eliminate and improve the tags was to look at whether or not they were 'cognitively basic'. Rosch proposed that commonly used concepts cut the world up at a 'basic' level which maximises within category similarity and minimises between category resemblance (Rosch 1978). Uschold and Gruninger suggest that words that are 'cognitively basic', in that they are couched at a common level of abstraction, are better candidates for use in taxonomies and as metadata (Uschold & Gruninger 1996). This is because it is at the basic level, 'dog' instead of 'mammal' for instance, where meaning is common to most people. Three frontier users were asked to indicate (using either 'yes', 'no' or 'do not know') whether the wording of each tag was at the level of abstraction generally used at Sainsbury's. These ratings were used to identify tags that could be considered for elimination or rewording.

The tags were also screened for relevance. In the case of the first set of documents this was done by examining the tags while bearing in mind user needs (as expressed by a document author). For the second set of documents, results from the information needs interviews were used. The interviews yielded a set of queries or questions which needed to be expressible using the tags. They also produced a set of rationales behind relevance judgements. Again, the tags could be tested to see whether they would provide the information required to discriminate effectively. This procedure allowed redundant tags,

or types of tag, to be excluded, while also identifying what types of tag might be usefully added.

A final rationalisation procedure was used to prune away the least useful of the large number of rarely occurring tags. The number of occurrences of each tag in all tagged documents was counted up. All the tags which only occurred one or two times were listed alongside their applicability. Tags were removed when they were highly applicable (i.e. applicability = 3) and their paragraphs could be assigned a more general existing tag (i.e. fruit -> perishables). Tags were also removed when they only occurred once, and an equally discriminating but more frequent alternative could be found. This pragmatic procedure allowed the removal of 49 out of a total of 238 tags, roughly 20%.

2.3 Metadata Taxonomy Design and Production

The structure, or taxonomy, that houses the metadata is crucial. Tags presented as a structureless mass would be unwieldy, if not unusable, by the document reader and metadata author. A taxonomy allows tags to be more rapidly located since it delimits a small subset of possibly germane tags. A taxonomy also allows the tags to be more easily and consistently authored, since parts of it (e.g. a class like 'competitors') suggest appropriate tags (e.g. the instance 'Tesco Supermarkets'). As such the taxonomy can be used to embody the typical information needs of a document's users and prompt the production of metadata which is aligned to them. This is important, since has been established (chapter 3 section 3.2.4) that authors cannot be relied upon to predict their reader's information needs. Hence, a taxonomy can be used to guide an author in the creation of the metadata needed by a document's audience.

The taxonomies developed share many features with several types of ontology used in knowledge based systems (see chapter 3 2.2). Consequently, techniques developed for creating these ontologies were used here. The taxonomies constructed in this research

are not formal ontologies using formal artificial languages, such as TOVE (Gruninger & Fox 1995) and $(KA)^2$ (Benjamins et al. 1999). However they do have a similar restricted and structured form of language to the structured informal ontologies which serve as the basis for formal ontologies (Uschold 1996). The taxonomies created here have more substantial similarities to the type of ontology defined by Jokela et al. (Jokela, Turpeinen & Sulonen 2000). This type of ontology is made up of a number of independent 'dimensions' (e.g. 'Business Area', 'Publication') each containing an internal structure (e.g. for 'Business Area', a hierarchical structure, and for 'Publication', a list structure).

The following is a description of the methods used to construct the two taxonomies produced during the course of this research. First we examine the methods used to construct the taxonomy for the first set of documents, then we go on to describe how these were improved for the second of documents.

The construction of the first taxonomy largely followed the methodology proposed by Uschold and Gruninger (Uschold & Gruninger 1996). The first stage consisted of the tag generation as described above. The taxonomy development was initiated by identifying natural semantic groupings of tags. To accomplish this, tags, each on a separate slip of paper, were moved around the surface of a large desk, grouped and regrouped, until a satisfying grouping had been found. By naming these groups, the first version of the taxonomy was produced. These names were inspected and refined for clarity and parsimony by both the researcher and frontier users. Many variations on the structure were considered in the search for something that was both parsimonious and concrete enough to be accessible to company employees. Once a structure had been settled upon, it was used in the tagging of some documents. These documents required some new tags, which themselves prompted some minor changes to the taxonomy. To summarize, the tags were first generated and semantically grouped, a taxonomy was

created to describe these groupings, and it was then iteratively refined through inspection and use.

As with the methods initially used for tag generation, the methods described above do not focus on producing a taxonomy to support goal-directed search explicitly. They support the construction of a reasonable taxonomy, but one that is organized solely around the structure of the domain. This structure is not tuned to the task of discriminating relevant content. The second taxonomy was developed with an improved method that used the information needs interviews described above to produce a taxonomy organized around the structure of the domain and along the dimensions used to make relevance discriminations.

The improvements to the taxonomy construction method were derived from the information needs interviews with members of the documents' audience. A first draft of the taxonomy was generated by grouping rationales according to the types of information they used. For example, the decision to reject a paragraph on the basis "I already know about the picking centre in Manchester" uses information about a business unit, i.e. 'picking centre' and a location, i.e. 'Manchester'. The grouping process allowed the common types of information used in decisions about relevance to be identified. A first draft of the taxonomy was constructed from these information types. Constructing the first draft in this way ensured that the taxonomy's main elements and structure supported the task of discriminating relevant paragraphs.

In order to ensure the taxonomy's completeness and parsimony, it was then tested and refined using Russell et al.'s learning loop process (Russell et al. 1993). This involved testing the taxonomy by applying it to the existing tag set. The tags that were not well housed, and the parts of the taxonomy that were not much used, motivated improvements. The improved taxonomy was then re-applied to the tag set, the taxonomy re-adjusted, the tag set re-applied and so on until a satisfactory fit was found.

The learning loop process provided a much more methodical and principled approach than relying upon inspection as recommended by Uschold and Gruninger (Uschold & Gruninger 1996). Once this had been done, the frontier users inspected the taxonomy's wording for clarity.

The second taxonomy was developed for a very different type of document from the first, but for the same community of users with the same 'industry' concerns and concepts. It therefore should have been possible to use the previous ontology as a starting point, but interestingly, the two taxonomies ended up looking very different. This suggests that these are not generic taxonomies that can be blindly applied to new document types; taxonomies of this kind need to be tightly coupled to the documents to which they must relate. This point, and issues surrounding it are discussed further in Chapter 8 section 2.2.

3 Visualization Design

Once high-quality paragraph-level document metadata is available, the next challenge is to provide a visualization that puts the metadata in the hands of the user. Here the GridVis visualization is described; first the design methods used are discussed, then a brief overview of the core features is given, followed by an illustrative scenario. Once the basics of the GridVis design have been made clear, the more advanced functionality and enhancements, and the rationales behind them, are addressed.

3.1 Design Methods

The design work was supported by a number of techniques: the close involvement of potential users, iterative prototyping, informal user testing and experimental evaluation. These techniques are only briefly described here. The results that impacted on the design of GridVis will be discussed in the next section.

Initial design ideas were refined and followed up by quickly assembled simple prototypes. The initial ideas were worked out on paper using general design and usability principles as guidelines (Shneiderman 1998). These were then honed, in part through discussions with frontier users, to make way for some simple implementations. The concretisation of ideas and sketches in interactive mock-ups was itself helpful in spotting pitfalls and design trade-offs.

These prototypes were then used for the user trial (chapter 5, section 1). The frontier users were asked to do various reading tasks using the prototypes. This allowed interaction problems to be uncovered and exposed limitations of the design. The initial few iterations provoked substantial changes to the design; later, when the foundations of the GridVis design settled, informal user testing informed more localized changes and additions.

A quantitative empirical evaluation of an interactive implementation of GridVis was also carried out (Chapter 5, section 3). Careful qualitative analysis of the data from this experiment enabled some subtle interface design problems to be identified.

3.2 The Visualization Design

3.2.1 Overview

The GridVis client-side application contains a visualization of a document's metadata (see Figure 27). This visualization can be considered as three interlinked sections: the metadata tree on the left-hand side, the iconic document overview running along the top, and the grid sitting at the centre. The taxonomy is used to create a tree structure and the metadata tags are added to this as leaf nodes. Each tag has a corresponding row in the central grid. The iconic document overview is a miniature image of the document laid out horizontally along the top of the visualization. Each paragraph in this overview lines up with one of the columns in the grid below. The central grid shows which metadata tags apply to which paragraphs in the document: each column represents a paragraph and each row represents a tag. Hence, to show that a tag has been applied to a particular paragraph, the cell where the appropriate column and row meet is shaded. The colour of the cell is determined by the value of the applicability attribute given to the tag when it was applied to a paragraph: the higher the applicability the darker the colour (a legend for this scale is always visible in the top left-hand corner). The metadata tags and taxonomy, iconic document overview and grid, are thus brought together to produce a visual overview of the document.

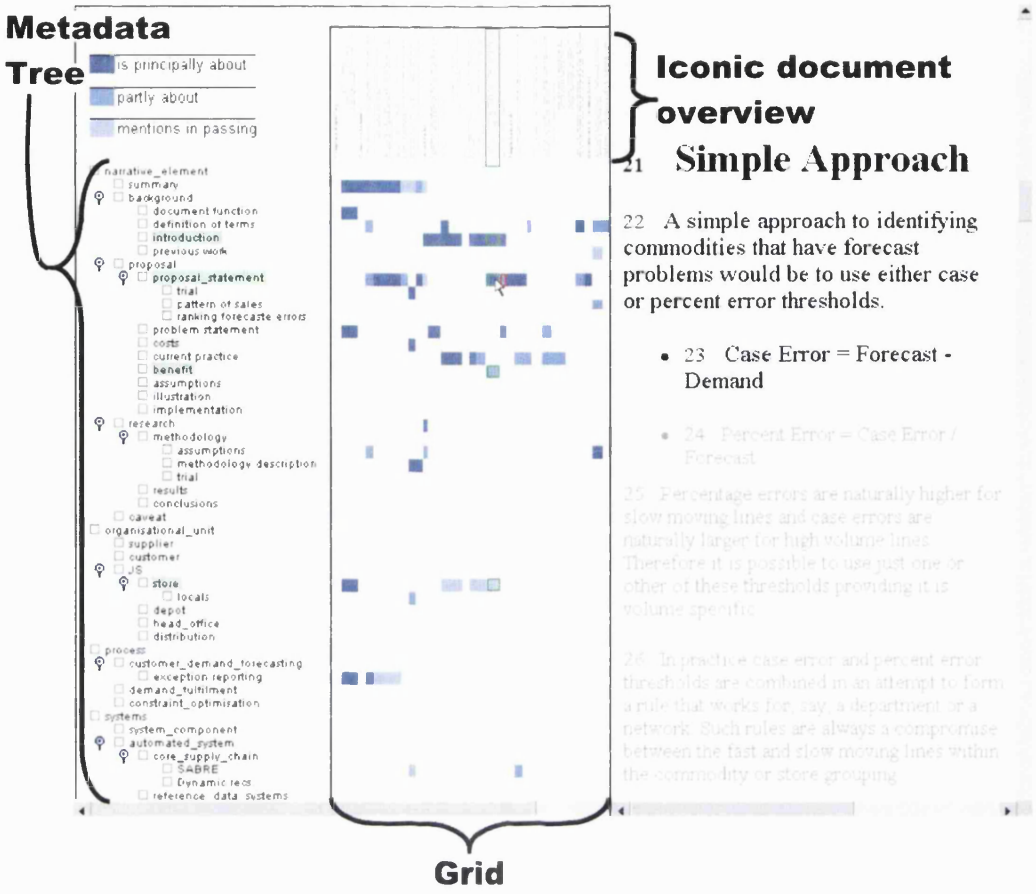


Figure 23 The components of the GridVis visualisation

The user can query the visualization with their mouse through a combination of dynamic querying and brushing (Card, Mackinlay & Shneiderman 1999). When the mouse is placed over a cell in the grid, all the tags in the column are highlighted; hence the paragraph's content is described. By looking along the row of an interesting tag, the user can see where and to what extent that topic comes up in the document. So by moving the mouse in the central grid, the user can explore what different parts of the document are about, and where in the document particular topics are covered.

The user can also query the document by clicking on any cell in the grid. Clicking on the grid will produce a query that is sent to the servlet formulated with the metadata and paragraph corresponding to the selected cell. The servlet will produce an HTML version of the document in which every paragraph tagged with the selected metadata is

highlighted using a bold font. The document window will display the paragraph selected, thereby offering details on demand (Card, Mackinlay & Shneiderman 1999).

3.2.2 An Illustrative Scenario

Scenarios (Carroll 2000) have been found to be effective design and communication tools. Consequently the scenario below is included to give a better idea of how the functionality described above might be used in practice.

Clair, a Sainsbury's employee, sits down to look at this week's 'industry update' newsletter, with a view to finding out whether the company plans to sell its US interests. She starts up GridVis and looks through the metadata to see if any of Sainsbury's US subsidiaries are mentioned. She scans down the top-level entries in the metadata taxonomy until she reaches 'business unit'. Then she scans the next level of entries under 'business unit' until she reaches 'Sainsbury's'. Under Sainsbury's she sees an entry for 'Shaws' which is a US subsidiary of Sainsbury's.

Having found a relevant tag, Clair proceeds to see if any of the paragraphs relevant to this tag will say anything about plans for a sell-off. She looks along the tag's row in the grid and sees that two cells have been coloured in. Passing her mouse over the first she sees a coloured cell in the same column and its tag, become highlighted. This cell's tag label reads 'loyalty cards'. Deciding that therefore the paragraph is not relevant, she moves her mouse on to the next coloured-in cell. In this case the 'acquisitions' and 'USA' tags light up. This surprises Clair, since she expected Shaws to be on sale, not making acquisitions; she decides to look at the paragraph and clicks on the cell. The right-hand window is refreshed; the paragraph she wants now appears at the top of the screen with its text in bold.

This scenario describes the simple case where the user is familiar with GridVis and the metadata taxonomy, and has a query for which there happens to be directly relevant

tags. A novice user might spend time exploring the capabilities of the tool only use some of the functionality. For example, they might decide on a tag without clicking on it, and investigate the paragraph cells by passing the mouse along the tag row. A user who has formulated a query for which there are no directly relevant tags is likely to follow a similar strategy. They would be forced to opt for the most relevant looking tag, or conclude that the document does not contain the information they need.

3.2.3 Design Rationale and Advanced Features

Descriptions of a design's end state misses out the many insights gained and put to work during its evolution. Since these insights are often transferable to related tools the more important ones are covered here.

The core design decisions behind GridVis were initially developed alongside alternatives that were cast aside on the basis of a task analysis as described in chapter 5, section 2. This task analysis made it possible to pick out key goals and objects involved in using paragraph-level metadata. The different designs being considered at that stage were compared in their ability to support these key goals. Each design provided some support for each goal but it was clear that the basic GridVis design would support them best.

3.2.3.1 Grid orientation

One insight gained from the user trial (chapter 5 section 1) was that subjects had trouble conceptualising the visualization's horizontal representation of the document in the iconic document overview. This problem underlined the importance of the iconic document overview in providing a representation of the document's orientation. In later versions, this function was bolstered by adding the ability to query the document directly from the iconic document overview; clicking on a paragraph in the overview

would highlight the paragraph's column in the grid and bring this paragraph up in the document window.

3.2.3.2 Navigation history

The user trial and experimental evaluation showed that people would often revisit the same paragraphs. Some subjects commented on the value of this; being led to the same parts of the text from different tags helped them see the text from different perspectives. Some subjects admitted that they had not expressly wanted to revisit paragraphs, but were simply going where the visualization led them. These problems led to the design of a history visualization feature. This uses colour markings on the grid to indicate which tags, cells and paragraphs had already been visited.

3.2.3.3 Paragraph context

Similarly, it became clear that users would want to read the section title of any paragraph they selected in GridVis. In the user testing sessions and experimental evaluation, people tended to scroll upwards from a paragraph they had been offered by GridVis, in order to read the section title. The section titles perhaps offer a way to contextualise paragraphs (e.g. 'method' vs. 'introduction'), thus making it easier to bring transferable knowledge about document structure to bear in the process of extracting relevant information.

The need to give paragraph context was resolved through several devices. Firstly, instead of displaying the selected paragraph at the top of the document window, the preceding paragraph was shown first, followed by the selected one. This proved very effective for documents with short sections such as news digest documents, since it often brought the titles into view. A second device, which was brought in later, was to

display the relevant section title above the 'iconic document overview' as the mouse rolls over columns in the grid (see Figure 26).

3.2.3.4 Making the grid scannable

Comments from the user trial suggested that users had difficulty in visually locating relevant cells; several features were added to make this easier. The essential problem here is that it is quite hard to look down a row or column in the grid to find coloured cells. This is because the eye can easily stray from the desired row or column. The first feature added, to deal with this problem, was a crosshair graphic. As the mouse was moved around in the grid, the borders of the column and row under the mouse would be drawn thicker and (in later versions) in a bright green colour. Although this helped, it forced users to bring their mouse into the correct row of the grid. This could be cumbersome, since the user's mouse movements over the grid have to be coordinated with the highlighting of tags in the metadata tree.

A further set of features were implemented to avoid this problem and deliver better support. This required the user simply to select a tag of interest rather then the more complex positioning of the mouse in the correct row of the grid. Once the tag is selected, the tag's row in the grid is highlighted with coloured borders and the columns of relevant paragraphs are given a different colour (see Figure 24). Giving the relevant cells a different colour is a particularly effective device. This is because it allows the cells to be located with an effortless pre-attentive visual search whose duration is independent of the number of cells in the display (Healey 1996).

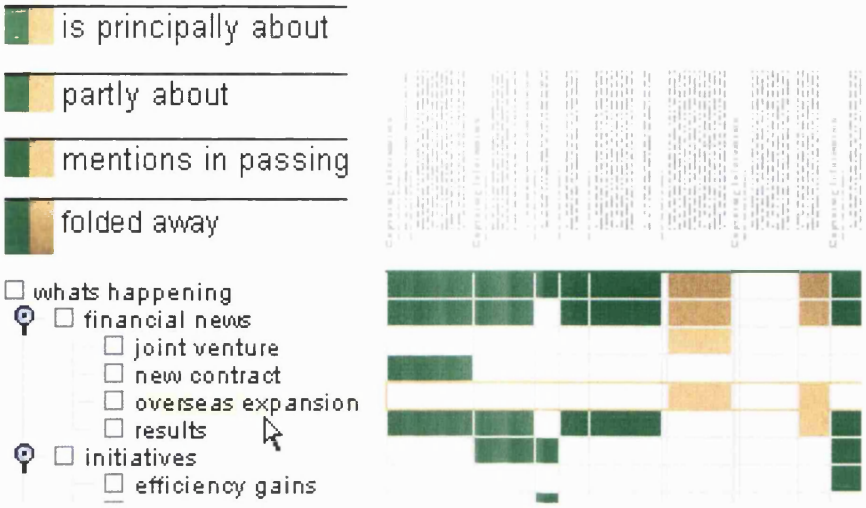


Figure 24 The paragraph cells relating to the tag selected are given a different highlight colour, allowing the user to locate them effortlessly.

3.2.3.5 Supporting the use of tags

Another important insight came from the task analysis based on data from the first round of user testing (chapter 5 section 2). This showed that the act of tag selection was a key part of using GridVis for goal-directed search (see Figure 25). Whereas the tags had been integrated into the grid in a mock-up interface (see Figure 34), subsequently they were displayed in a separate and conventional vertical tree display. This change served to more clearly display the tags and hierarchical organization, using a sophisticated yet familiar interface.

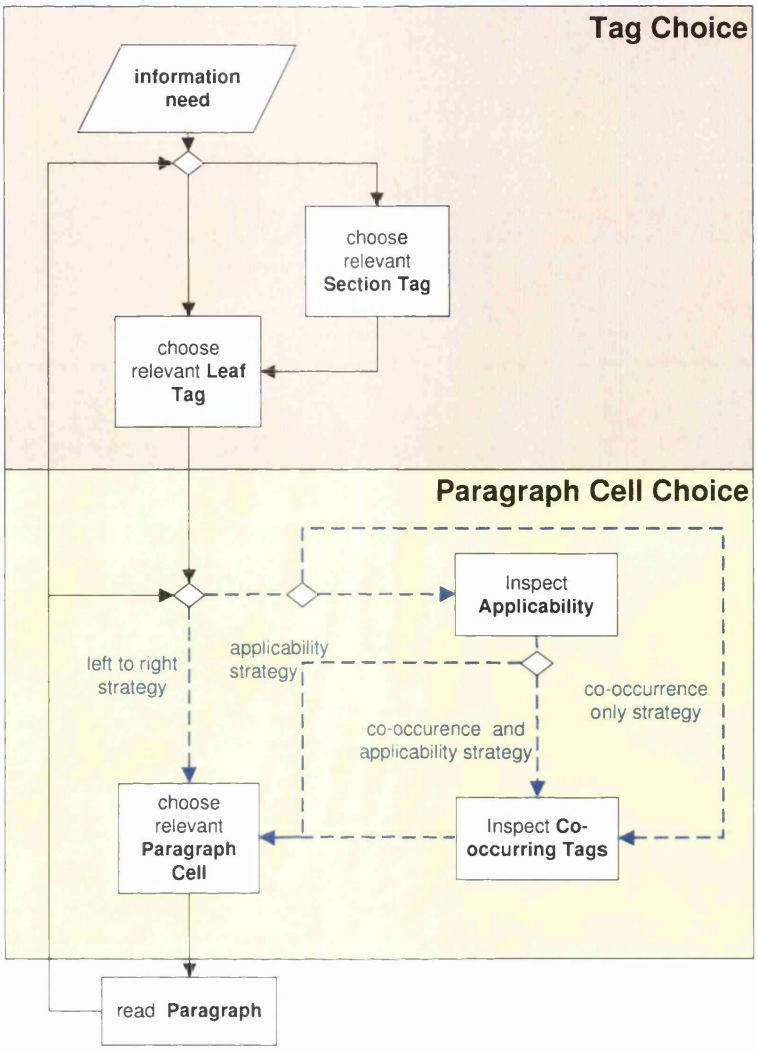


Figure 25 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold),the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines)

Once a tag is chosen, a paragraph (i.e. cells) relating to it must be selected. This task has several components. First potentially relevant cells must be located, then these must be inspected in order to decide on the relevance of these paragraphs to the query. The design decisions that went into making the cells easy to locate will be discussed first, and then improvements intended to make their relevance easier to assess are described.

Comments from the user trial and the frontier user group suggested it was hard to locate the relevant cells. Also, the task analysis (chapter 5 section 2) showed that this was a

key goal in GridVis use (see Figure 25). The qualitative analysis of data from an empirical evaluation of GridVis allowed this task to be looked at in more detail (chapter 5 section 3.3.1). It revealed that the task could be tackled using four different strategies (see Table 10 and Figure 25). It also indicated that subjects did not make much use of any co-occurrence strategies. This suggested that GridVis was not adequately supporting the use of co-occurrence information.

Strategy	Information used
Left to right	None
Applicability	Applicability of cells on tag of interest
Co-occurrence only	Relevance of other tags in columns (paragraphs) of interest
Co-occurrence and restricted applicability	Applicability of cells on tag of interest Relevance of other tags in columns (paragraphs) of interest

Table 10 Different possible strategies for paragraph cell selection

To gather co-occurrence information, the users must look at the other tags on the paragraph of interest. Subjective experience suggests that co-occurrence information is a powerful way of discriminating relevant from irrelevant material. Together with the tag of interest, it gives a reasonably full understanding of a paragraph's content. Users' tendency to not use this information could be attributed to the fact that it is difficult to do so; upon locating a cell they would have to look up the column to find other cells that are filled in, then look along the rows of these cells to read their tag labels.

To tackle this problem, all the tags on a paragraph are highlighted when the mouse passes over its column. This relieves the user of the need to look up co-occurring tags on the grid before reading them; they need only pass the mouse over the appropriate column and read off the highlighted tags. The investment required to use the co-occurrence is thus reduced, making it more attractive to the user.

A qualitative analysis of the later experimental evaluation revealed that when selecting paragraphs relating to a tag of interest, users still seldom used co-occurrence information. A further analysis found that precision could be substantially increased by making full use of the co-occurrence information. Hence despite the possible rewards, the cost of using co-occurrence information was still too high.

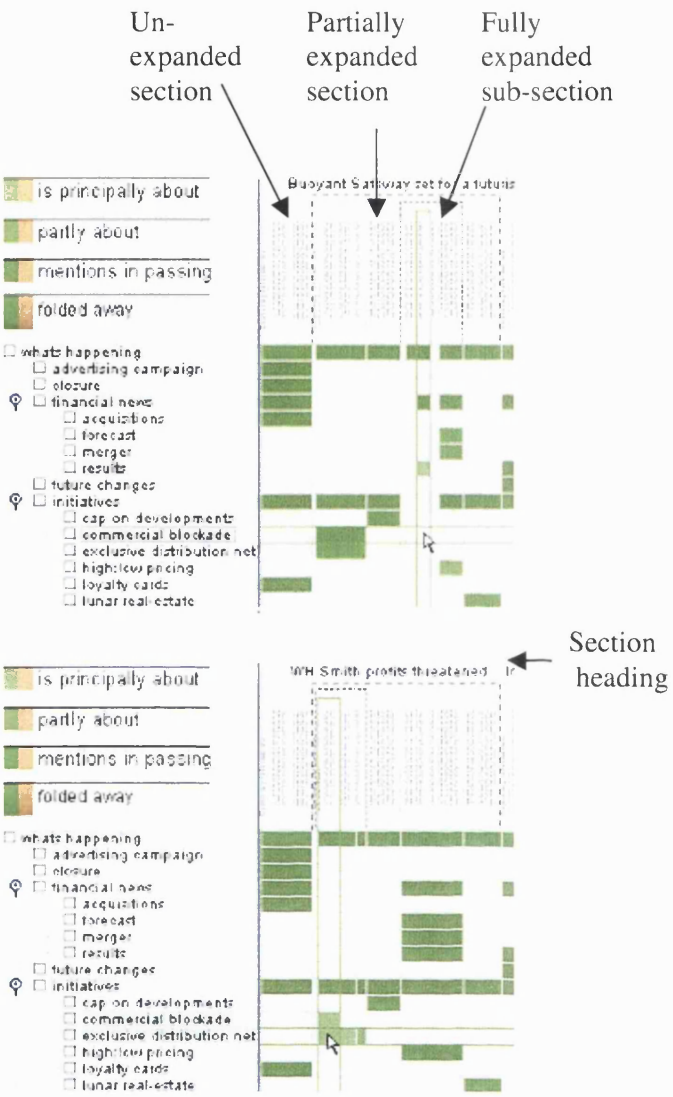


Figure 26 An illustration of the dynamic level of detail management feature

A further problem impeding the use of co-occurrence was later revealed by a GOMS analysis of cell selection (see chapter 6 section 2.1.1). The GOMS analysis essentially decomposes the process of selecting a cell using the co-occurrence strategy into component actions for which standard time estimates can be made using results published in the psychology literature. This analysis revealed that the reading of tags highlighted, for the paragraph of interest, was *the* major contributor to task time. One way to reduce this contribution would be if the user only read the leaf tags. It is the leaf tags that are of major importance; the parent tags serve

mainly to contextualise them. By only reading leaf tags, the GOMS model predicted that the task time for instance selection using co-occurrence would be reduced by between a third and half. To realize these gains, a simple change was made to the interface which allows the user to easily ignore parent tags. The highlight colour was made a light grey for parent tags, as opposed to the light green and orange given to child tags. Thus, via a GOMS analysis, a design problem was identified; this prompted a simple design change, which substantially sped up the use of co-occurrence information in cell selection.

New functionality has been added to GridVis, to further address the high cost of using co-occurrence. The co-occurrence strategy was made easier through partial automation; if the user selects tags of interest, GridVis indicates which tags co-occur on each paragraph. This can be used in several ways. Most simply, it can be used to see which paragraphs relate to the largest number of interesting tags. Perhaps more often it can be used to find a paragraph that has *all* the tags of interest applied to it.

Another possible factor inhibiting the use of applicability and co-occurrence information is GridVis' voracious appetite for screen real estate. With longer documents, the users have to scroll horizontally to see the whole grid. This makes it impossible to see every paragraph relevant to a tag and therefore confounds attempts to select the most appropriate paragraph.

The screen real estate problem was addressed by an implementation that dynamically redraws the grid at varying levels of detail to only show in full the parts of the document relevant to the user's current activity (see Figure 26).

4 Alternative design features considered

This section discusses a number of design features that were considered but ultimately rejected or left unimplemented.

As it is, GridVis does not give users the ability to change the arrangement of the metadata tree. There are a number of possibly useful ways in which a user might change the tree's arrangement. Perhaps the most obvious is to offer users the ability to fold away portions of the tree that are not of interest. Traditionally portions of the vertical indented folding tree visualisation used for displaying the metadata can be folded away. This option would be particularly useful in situations when there are too many metadata elements to fit on one screen. However, such an option might lead the user to unnecessarily invest time in deciding which branches were not of interest and in folding them away. Another possibility is to automatically prune the tree according to the currently selected or highlighted metadata items. This would mean that the metadata items selected and those that it co-occurs with would be shown together omitting all the other tags and their rows in the grid. This would reduce the number of rows in the grid, avoid the need to pick out highlighted metadata items and reduce the distance the eye has to travel. However, both this feature and the folding feature would lead to a serious problem. The changes to the tree and grid are likely to disorient the users. Once items are pruned or folded away the structure of the tree is changed and spatial memory cannot be relied upon to rapidly locate items of metadata. A better solution might be to use a fisheye view to reduce the size of metadata elements that are distant from the current foci as defined by the mouse and the currently highlight metadata items.

Another possibility would be to allow the user to use a keyword search to locate relevant metadata items. Instead of visually searching for the metadata terms of interest they could enter a search term into a simple search engine that would highlight related

or matching metadata terms. Although in some cases this may save the user time, it would lead to the metadata being under used. The difference between search terms and the metadata embedded in a taxonomy would mean that searches would often miss relevant metadata tags and find irrelevant ones. This is because the meaning of metadata often depends on its location in the taxonomy. For example the tag 'results' is fairly meaningless unless one knows it is within the 'financial news' section. Also search terms cannot easily be used to take advantage of the conceptual structure embodied in the taxonomy or the domain knowledge that resides in the user. For example, the search term 'abroad' will not pick up those items in the 'location' section, such as 'Malaysia' or 'USA', which are not in the UK. To summarise, the introduction of a keyword search is not likely to be a great aid, furthermore it may distract the user from paying the attention to the taxonomy needed to do a good job of locating relevant tags.

5 System Implementation

In this section the use of XML to encode the documents and taxonomy is described then an account of the visualisation's implementation is given.

5.1 Encoding the metadata in XML

The documents and their metadata and taxonomies discussed in section 2 needed to be encoded in a technically appropriate format, so that they could be conveniently authored and visualised. XML was chosen to do this since, as discussed in chapter 2 section 2.3.2.1, it has been designed to allow structured documents to be conveniently encoded and manipulated.

Each document's structure, text and metadata are encoded in a single XML file. An XML-Schema was used to ensure well formedness and to facilitate the metadata authoring process. The metadata was encoded alongside the text to simplify technical requirements. It allows an XSL-T style sheet to directly filter and format paragraphs according to the metadata assigned to them. Also, encoding the metadata this way allows GridVis to derive attributes from the text and the document structure (such as paragraph length and section level) for use in the visualisation.

The XML tags were designed to allow an arbitrary hierarchical document structure and enable rich metadata tagging at the paragraph level. The different XML tags used and their legal relations to each other are captured in Figure 27. The nested <collectionTag>s are used to encode document structure. A <terminalTag> holds a single paragraph whose text sits in its <text> element, and whose metadata sits in its <Description> element. The <Content_tag> contains an attribute which holds both the tag and its location in the taxonomy (e.g. 'publication.The_Independent'), and one which holds applicability.

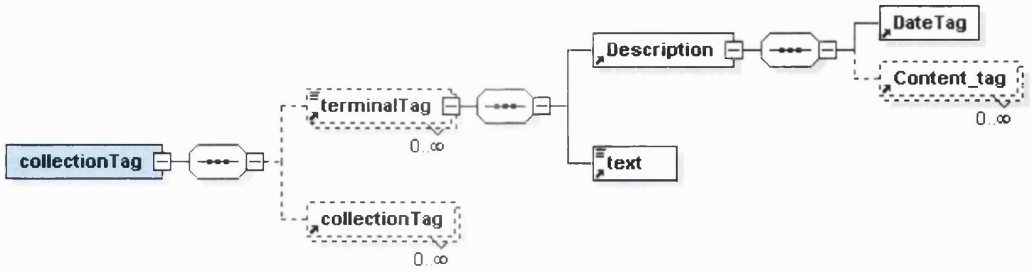


Figure 27 The structure of the XML documents visualised by GridVis. (Dotted lines indicate optional elements)

The XML standard for describing metadata is RDF (see chapter 2 section 2.3.2.1). This was initially used but proved unsuitable, since it has been designed to provide data for AI reasoning algorithms. Since no such algorithms are used in GridVis, RDF is of no immediate utility. Moreover, the use of RDF would have complicated the production of the metadata. Since it would require separate elements for applicability, it is only the <description> elements that follow the RDF standard.

An XML-Schema was not used to facilitate the authoring process for the first set of documents. When the first set of documents were produced, the XML-Schema standard had not been settled. The consequence of not using XML-Schema (or DTD's) were two-fold. It necessitated error-checking code in GridVis which could not provide very targeted debugging information; this in turn increased time spent debugging the XML. Another consequence was a tendency for equivalent tags to turn up in different places in the hierarchy, or with slightly different spellings (e.g. 'store' and 'stores'). These errors had to be eliminated by a drawn-out and tiresome process of manual inspection.

In order to avoid similar problems, an XML-Schema was used, alongside a specialised XML authoring environment, to facilitate the authoring process for the second set of documents. The Schema defines each allowed XML element, the attributes it may have, the types those attributes make take and the legal relationships between elements. It also

allows for all the possible values of an attribute to be defined. This feature was used to ensure that tags were consistent and correct across paragraphs and documents. The XML-Spy authoring environment used allowed the tags defined in the schema to be selected from a drop down menu. The use of XML-Schemas alongside a specialist XML authoring package, made the process of authoring the XML within-document metadata much faster and easier.

5.2 The visualisation implementation

GridVis is designed for use on a corporate intranet where software installation for low priority applications is problematic, it therefore has a client-server architecture; the visualization is generated by a client-side Java applet, the queries are answered by a Java servlet using XLS-T to produce customized HTML documents.

GridVis starts by building the data structures, needed for the visualisation, from two XML documents. One document defines the metadata taxonomy, and the other contains document content and accompanying metadata. The taxonomy is used to build the metadata tree (see Figure 23). The metadata for each paragraph is read off and assigned to the appropriate section of the metadata tree. Each paragraph and section is represented by a *Chunk* object, which retains the paragraphs size, section heading and HTML anchor number. These *Chunk* and *Tag* objects are stored in both linear (*Vector*) and hierarchical (*DOMTree*) data structures. The applicability of each piece of metadata to a particular paragraph is stored in a, *Chunk* by *Tag*, matrix (i.e. a two dimensional array). For the dynamic level of detail management feature, other data structures must be built; these will be referred to as the visibleGrid data structures. These essentially mirror, and are built from, the data structures just described, but only represent the currently visible data.

GridVis uses these data structures to build the visualisation. The matrix of applicability information is used to construct the grid, while the linear data *Chunk* data structure is used to produce the iconic document overview, and the hierarchical *Tag* data structure is used to build the metadata tree. If the level-of-detail management feature is being used, each user event results in the visibleGrid data structures being rebuilt using simple recursive tree traversal algorithms.

6 Chapter summary

This chapter has described the design and construction of a tool for using paragraph level metadata to support the goal-directed search of corporate documents. This involved the careful production of high-quality metadata as well as the construction of taxonomies to house them. The metadata and taxonomies were constructed using methods from the formal ontology construction literature, as well as new methods based upon interviews with potential users, focusing on how they would distinguish the material they would read from that they wouldn't. The GridVis tool was developed and refined through several rounds of evaluation and analysis. The rationales for design enhancements are discussed, drawing on evidence from user testing, empirical evaluation, and various analyses. The next chapter describes initial evaluation work, which includes users testing, task analysis and a controlled experiment.

Chapter 5 Initial evaluation

The previous chapter described the process used to construct the paragraph-level metadata and the design of GridVis. This chapter describes three pieces of evaluative work carried out in the process of developing GridVis. The first, the user study, is a small formative experiment assessing the potential of paragraph level metadata for supporting three different types of reading task. The second, the task analysis, uses data from the user study to produce a detail task model of the use of GridVis for goal-directed search. The third, the experimental evaluation, evaluates a fully interactive prototype of GridVis.

1 User trial – Exploring the capabilities of GridVis

A small user trial was carried out which investigated how and to what degree GridVis might support goal-directed search. This provided useful insights into the shortcomings of the GridVis design. The study also provided data for use in the task analysis, reported later in this chapter (see section 2), on how GridVis is used for goal-directed search. Finally, it highlighted methodological concerns, which informed the design of the experiment reported in the final part of this chapter (see section 3.1.5).

This study was conceived to scope the capabilities of GridVis, and thus examined three separate reading tasks: overview, sorting, and goal-directed search. For the purposes of this thesis it is only necessary to consider detailed results for the goal-directed search task. However, in order to interpret these results it will also be necessary to examine the result for the overview task. The four participants were given the three different tasks to complete using GridVis. The performance on the goal-directed search task is analysed

alongside data from debrief interviews. The results suggest that GridVis might support goal-directed search.

1.1 Method

All of the four participants were frontier users. They were experienced knowledge workers and as members of the team that produced the document used in this trial, they were familiar with its subject matter. The experiment was carried out in a quiet room at Sainsbury's.

The three reading tasks set, getting an overview, goal-directed search and sorting, were taken from a taxonomy of work related reading types (Adler et al. 1998). The overview task relates to Adler et al's Browsing, here the reader was asked to "obtain a general sense of [...] content and extent" (Adler et al. 1998, p. 244) by simply inspecting the visualisation. For the goal-directed search task, which relates to Adler et al's 'Searching to answer questions', participants were asked to use GridVis to locate information relevant to one of nine questions. In the sorting task, subjects were asked to inspect GridVis, and rate the relevance of the document to different people.

The same document was used for all of the three tasks. This was an analytic research report, looking at the effectiveness of a set of supply chain rules, written by a fifth frontier user. The author of this report also provided the supplementary materials needed for the tasks, such as the questions for the goal-directed search task, and descriptions of potential readers for the sorting task. An effort was made to ensure that the materials used in this experiment were ecologically valid. The questions were deemed, by the author, to be the kinds of questions that readers of the document might ask.

The early mock-up of GridVis comprised a static image, an applet which sent the queries and a small document window (see Figure 28). The image was created in Excel, a simple java applet was written to turn selections into appropriate queries. The query was sent to an early version of the document filtering servlet. The servlet returned the document in a small browser window (see Figure 29), so that only the paragraph selected could be viewed. This small browser window was used to force the participants to rely on GridVis for navigation.

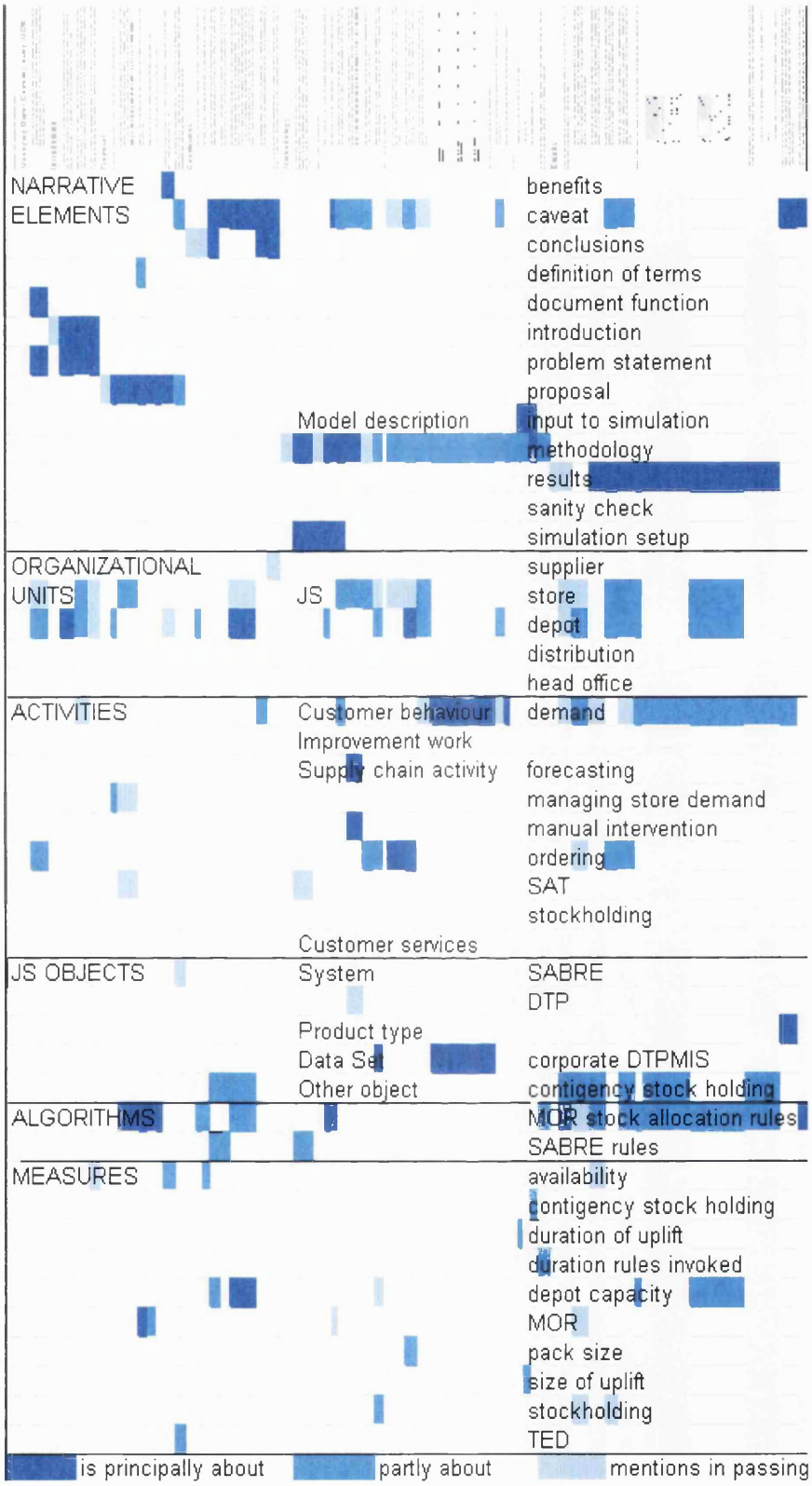


Figure 28 Screen shot of prototype GridVis mock-up used

The experiment started with a familiarisation session in which the use of GridVis was explained and each participant was encouraged to explore a document using GridVis. Once a participant claimed to be comfortable with GridVis, the experiment was begun.

For the first task, the overview task, a different document was shown in GridVis, just as depicted in Figure 29, except that a small Word document was shown in place of the query and document windows. This Word document contained instructions for the task and a place for the participants to write their summary of the document's content.

Once the participant felt satisfied with their summary, they moved on to the sorting task using a hypertext link to bring up the next screen. On this screen the same document was visualised in GridVis, but the accompanying Word document, contained instructions for the sorting task, and a form for the answers. Each participant was instructed to rate the documents' relevance to each of six readers.

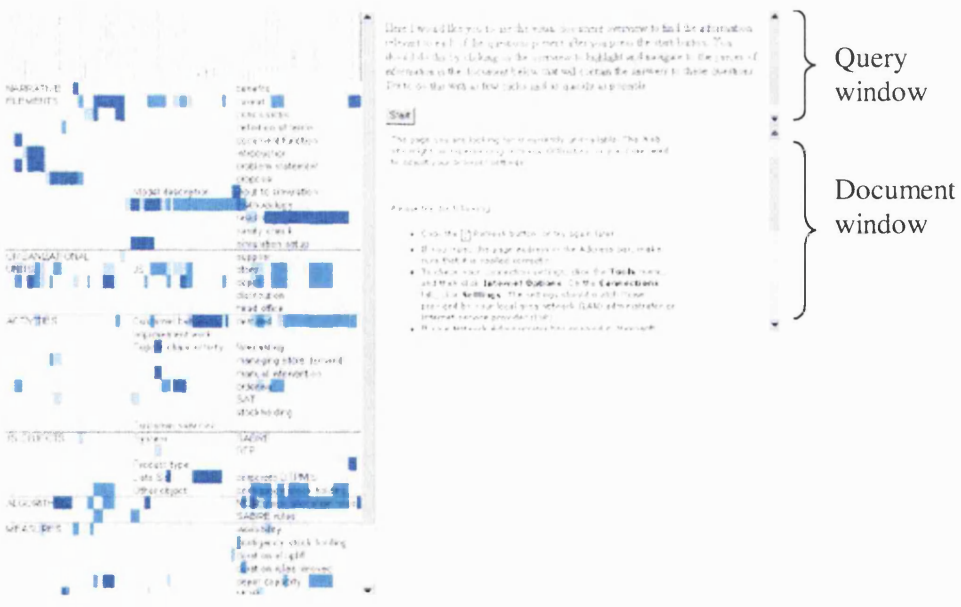


Figure 29 Evaluation interface for the goal-directed search task

The participants were then moved on to the goal-directed search task. Here they were presented with the interface shown in Figure 29. Once they had read the instructions they selected the start button and were presented with the first of nine questions. The participants were encouraged to solely rely on GridVis, and only use the scroll bar when the selected paragraph was partially out of view. They were asked to select a button labelled 'got one' each time they found a relevant paragraph. This allowed for the recording of the time it took to locate each relevant paragraph. Once the participant was satisfied that they had found all the relevant information available in the document they pressed the 'next question' button to move on.

The next phase of the experiment was designed to collect data on the quality of responses given for the first two tasks. The participants were presented with the browsing and sorting tasks using the same document, but using a full Internet browser window without GridVis. The participants were asked improve their answers to the overview and sorting tasks, using the full document and taking as much time as they needed.

Once this was done, the participants were briefly interviewed to elicit any comments about usability problems and the utility of GridVis for each task. Each interview was recorded.

1.1.1 Data analysis

The data from this experiment comprised of a log of every paragraph cell selection, as automatically recorded by GridVis, the answers typed into Word documents by participants, and notes on the debrief interview. The data for each task were analysed separately.

For the overview task, the analysis compared the summaries written using the GridVis visualisation and the document text. The summary written using GridVis was scored against the one written with access to the full document text. Each summary was broken down into the distinct propositions made and the objects mentioned. This allowed improvement made to the summary upon seeing the document text to be quantified in terms of the number of propositions omitted, added, refined and corrected, and the number of objects in the GridVis summaries reused in the document-based summaries.

In order to measure retrieval performance, the interactive recall and precision metrics, described in chapter 2 section 5.1.2.1, were employed.

1.2 Results

Only the results for the goal-directed search task will be discussed here; the results for the overview and sorting task are presented in Appendix 1.

The mean interactive recall, across participants and questions was 0.33, the mean interactive precision was 0.32. These figures are however somewhat misleading since there were no paragraphs relevant to two of the nine questions. This had the effect of elevating the scores, since when a participant successfully failed to find an answer in the text, for these questions, they were given a interactive recall of 1. The interactive recall and interactive precision scores for the remaining seven questions, for which relevant paragraphs did exist, were 0.18, and 0.33 respectively. These scores denote poor performance; they imply that only a fifth of the relevant material is being found, while a quarter of those identified as relevant in fact are not. However, as with the results from the sorting task, this rather poor performance might be partly attributable to differences in interpretations of questions by the document author and the participants in the study. Nevertheless, one participant did not find *any* paragraphs he thought were relevant,

despite several relevant paragraphs being present. This result suggests that participants genuinely did have problems finding relevant information using GridVis.

Observation of the participants using GridVis and the debrief interviews provided a rich source of information on usability issues. Some of these served to confirm the results given above; several participants said they found it hard to write meaningful summaries using GridVis alone. Many of the other comments were used in designing the next version of GridVis; the most pertinent of these were as follows:

- Two participants reported that the structure of the metadata tree/taxonomy was difficult to follow, which made finding tags difficult.
- Three of the four users took some time to start using co-occurring tags in deciding between paragraph cells, and one reported that doing so was too difficult. Another user said she found it hard to visually follow the grid's columns.
- Two participants reported that they often found themselves unintentionally returning to the same paragraphs. Inspection of the paragraph cell selection log showed that this occurred for all participants.
- One participant felt that GridVis was good for answering questions that were specific and had some relation to the tags.

The data collected in the observation of the participants using GridVis and the debrief interview that followed were subsequently used for the task analysis described later in this chapter (see section 2).

1.3 Discussion

Performance on the goal-directed search task was poor; participants achieved low interactive recall and interactive precision. Some of this low performance can be

ascribed to a number of methodological problems; only one judge was used to create the relevance judgements against which performance was scored, and participants were not given the chance to get an overview of the documents content before searching. However much of it can be attributed to problems in the GridVis interface.

The use of only one judge meant that interpersonal differences in interpretation (of the questions, or document text) had a large effect on performance scores. Substantial variation in relevance judgements between judges was observed during the experimental evaluation described later in this chapter (see section 3.1.4). The use of several judges allows a variety of opinions to be captured in the final relevance judgements used for score. This can mitigate the effect of interpersonal differences in interpretation. In this study however, these differences in opinion were treated as errors. Nevertheless, the inability of participants to find any answers to some questions suggests that the poor performance was not purely an artefact.

The other methodological factor that might explain poor performance was the fact that participants did not have any overview of the document they were searching. The participants did not read the management summary. Additionally, the very poor performance on the overview task (see Appendix 1) shows that GridVis does not give an effective overview of document content. This lack of overview makes interpreting and answering the questions difficult. For example, knowing that the document is a proposal to ensure that depot capacity is not exceeded, would help determine which parts, and how much, of a document might answer the question: 'what governs depot capacity?'. Hence, it is possible that performance could be improved if participants were allowed to read the management summary before attempting goal-directed searches. Another set of possible causes for the low performance relate to the inadequacies in the design of this primitive GridVis prototype; these are discussed further below.

Some of the shortcomings in GridVis's support for the goal-directed search task were uncovered in the debriefing interviews. Overlaying metadata labels on the grid made some of the tags hard to read (see Figure 28) and meant that the hierarchical organisation was not obvious. Also some of the information in GridVis might not have been accessible enough. For example, co-occurring tags can be used in deciding between paragraph cells, but to do so required considerable effort; the user had to visually pick out the other filled-in cell on the column of interest, then look along each row to read the co-occurring tags. Another point coming out of the interviews was that users were unable to determine what they had inspected and consequently found themselves unintentionally revisiting the same paragraphs. It is possible that goal directed search performance with GridVis would be much improved if these, and other, problems were addressed by changes to the design of GridVis.

Whether or not GridVis can offer good support for goal-directed search was not established in this study. The low performance measured is consistent with, although perhaps not entirely due to, a number of extraneous factors. GridVis's true performance was obscured by methodological problems which needed to be corrected in future evaluations. This study also exposed a number of usability problems whose resolution should improve performance. Finally, observation of the tool in use, along with interview and log file data, provide a rich base from which to do the task analysis described in section 2. This task analysis provided a thorough understanding of GridVis's use in goal-directed search. This understanding is needed to review the design effectively and realise the potential of the use of paragraph-level metadata visualisation for goal-directed search. The improved design was then evaluated in the experimental evaluation described later in this chapter (section 3), taking into account the methodological lessons learned in this study.

2 Task analysis

The user trial revealed some problems with the support of goal-directed search by GridVis. It was therefore important to carefully review the design if the true potential of paragraph-level metadata was to be brought out in future prototypes of GridVis. In order to best review the design, a task analysis was undertaken examining the task of *using paragraph level metadata for goal directed search.* This task analysis was grounded in data from the user trial described above. The analysis was however focused on the goals, actions, and objects inherent in the use of paragraph-level metadata for goal-directed search. The resulting task analysis helped direct design and evaluation by clearly articulating the procedures of a task.

Johnson's knowledge based task analysis (Johnson 1992) was the particular technique used since it was deemed most useful for focusing the design effort. It requires a taxonomy of the objects and actions inherent in the task, and an articulation of the task procedures. Inspection of the resulting model allows the isolation of key goals and objects for which the design must be optimised. Alternative designs for these key objects can be generated and their relative level of support for the key actions can be analysed.

The task analysis was carried out using data from two evaluations. It was initially based on the observational and interview data from the user trial already described above in section 1. It was subsequently refined, using the results of a qualitative analysis, reported later in section 3.3.1, to isolate the different strategies for tag and cell selection in the experimental evaluation.

The full result of the task analysis is a list of the objects, actions and procedures involved in using GridVis and a task model which captures their interrelationships. The

actions and objects are listed in Table 11. The task model is captured in the object taxonomy (see Figure 30), the goal structure and the procedural substructures have been depicted in a single flow diagram (Figure 31) so as to better convey their inter-relationships. A simplified flow diagram of the task model (Figure 32) brings out the basic features of the task.

Objects	Actions
Information need	Deduce Relevant Content Areas
Relevant content areas	Deduce Relevant Keywords
Relevant keywords	Deduce relevant Tag^{12}
Document	Evaluate relevance of Tag^{12}
Tag	Evaluate match of Tag^{12} Word to Keyword
Section tag	Evaluate match of Tag^{12} Word to Content Area
Leaf tag	Evaluate match of Tag^{12} Word to Information Need
Co-occurring tag	Select Tag^{12}
Word	Select Paragraph Cell
Paragraph	Read Paragraph
Paragraph cell	Evaluate Relevance of Paragraph
Applicability	Evaluate Paragraph Cell Applicability

Table 11 The objects and actions uncovered in the task analysis

2.1 Objects and the object taxonomy

The objects listed in Table 11 are best read in conjunction with the object taxonomy shown in Figure 30. The different objects uncovered by the task analysis and their relationships to each other will now briefly be described.

The 'Information need', can lie between the first three levels of Taylor's four level spectrum of 'question formation', discussed in chapter 2 section 4.1 (Taylor 1968); between the visceral need (actual but unexpressed), conscious and formalized questions such as those supplied in the user trial described in section 1 above. Information needs can be transformed into more formalised needs, such as *relevant content areas* or into

compromised needs, Taylor's fourth level, such as *relevant keywords* that might exist as *tags*. So for example the query "what is Tesco doing overseas?" might be transformed into a *relevant keyword* such as 'Tesco' for which a *tag* is likely to exist, and a *relevant content area* such as 'location' to which many *tags* might relate.

The *document* object contains metadata as well as the customary *paragraph*. The primary metadata object is the *tag*, which is specialised into three sub types; *leaf tag*, *section tag* and *co-occurring tag*. All *tag* objects have a human readable label made up of *words*. *Leaf tags* describe *paragraphs*, and are organised in a taxonomy of *section tags*. *Section tags* can take on the *paragraphs* described by their *leaf tags* and can also be used to directly describe *paragraphs*. *Co-occurring tags* are the *leaf tags* or *section tags* which occur on a particular *paragraph*. The *applicability* object records the degree to which a *tag* describes a *paragraph*. It can be inspected in a *paragraph cell*, which is the intersection of a *paragraph* and a *tag*.

---

[12] Here 'tag' an be substituted for 'Leaf tag', 'Section tag', or 'Co-occurring tag'

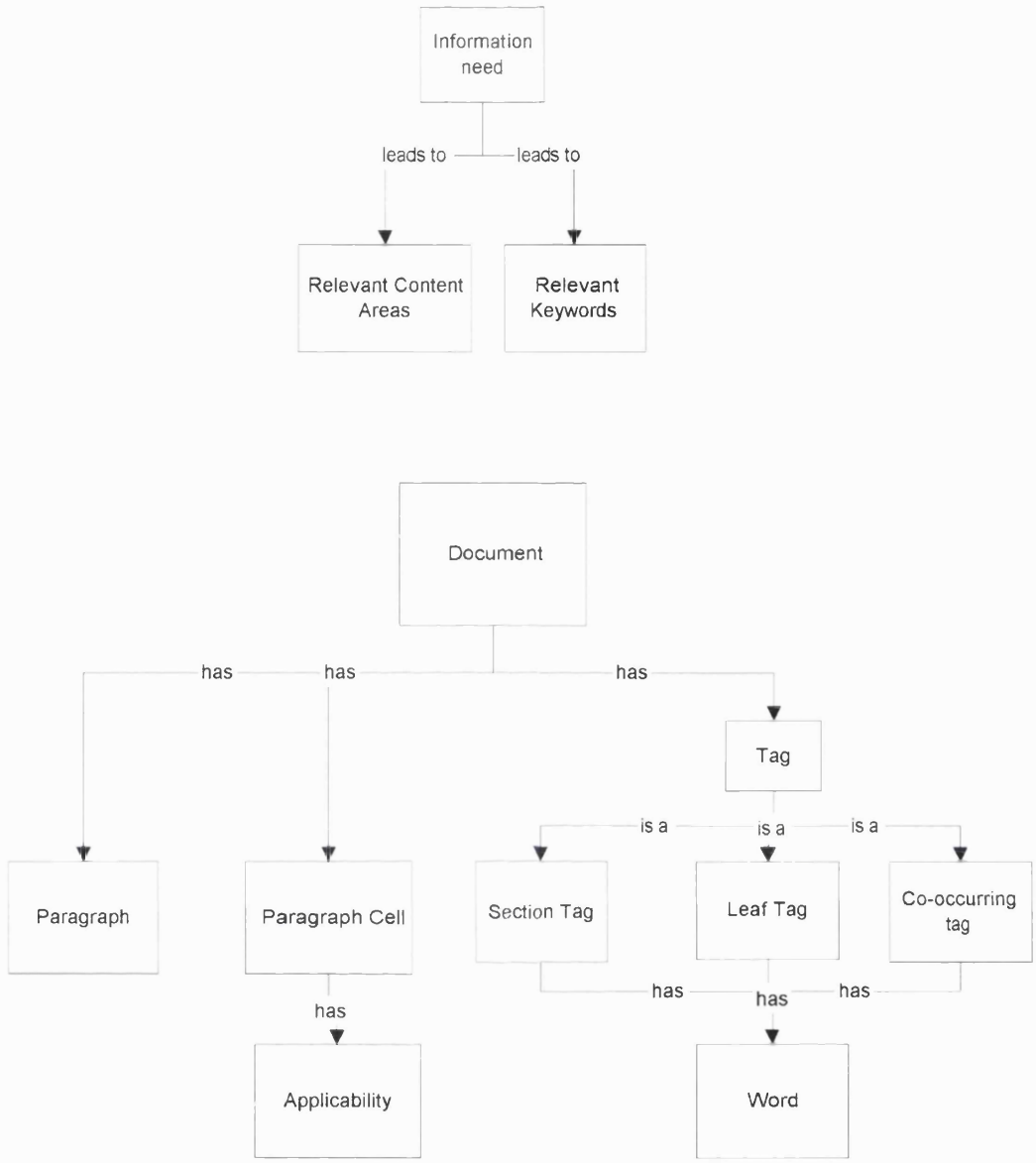


Figure 30 The object taxonomy

2.2 Actions, the goal structure and the procedural substructure

The actions listed in Table 11 are best read in conjunction with Figure 31 which shows the goal structure and the procedural substructure. These actions and their relationships to each other and to the objects just described will now be discussed.

The process of using paragraph-level metadata starts with an *information need.* This *information need* can be formalised via the deduce relevant content areas or deduce relevant keywords actions.

The *information need* is then used to locate a pertinent *tag* using a number of alternative approaches. One way of doing this is to first identify a relevant *section tag*, and then identify the relevant *leaf tag*s within that section. Another is to ignore the taxonomic structure and go through all the *tag*s (i.e. both the *leaf tag*s and the *section tag*s). When the user is not searching for a particular *tag*, they can evaluate the match of the *tag* to the *information need.* The *information need* may have been transformed into various formalisations, such as relevant keywords and relevant content areas. The user may therefore evaluate the match of the *tag* to each of these formalisations as well.

Once an appropriate *tag* has been found, the user can either select the *tag* or go straight on to locating an appropriate *paragraph cell.* By actually selecting a tag, the user may be able to trigger some changes to the visualisation which supports the search for *paragraph cell*s. As with locating *tag*s, there are several strategies for locating the most relevant paragraph cell on a tag. The simplest is to locate the next paragraph cell, where the user locates the first, unselected, *paragraph cell* with non-zero *applicability*. The user can also locate paragraph cell on tag with highest applicability by reading the *applicability* of each *paragraph cell* and selecting the most applicable. Finally, they can locate most relevant paragraph cell using both the applicability information and the co-occurrence information. This involves evaluating the relevance of each paragraph cell, by reading off its *applicability*, and reading and evaluating the relevance of each *co-occurring tag*.

Once an appropriate *paragraph cell* is located, it is selected, and the paragraph it leads to is inspected. If it turns out to be irrelevant, it is abandoned, and the user goes back to

look for another pertinent *paragraph cell* on the *tag* last investigated. If there are none, then the user goes back to looking for another relevant *tag* with which to continue the search.

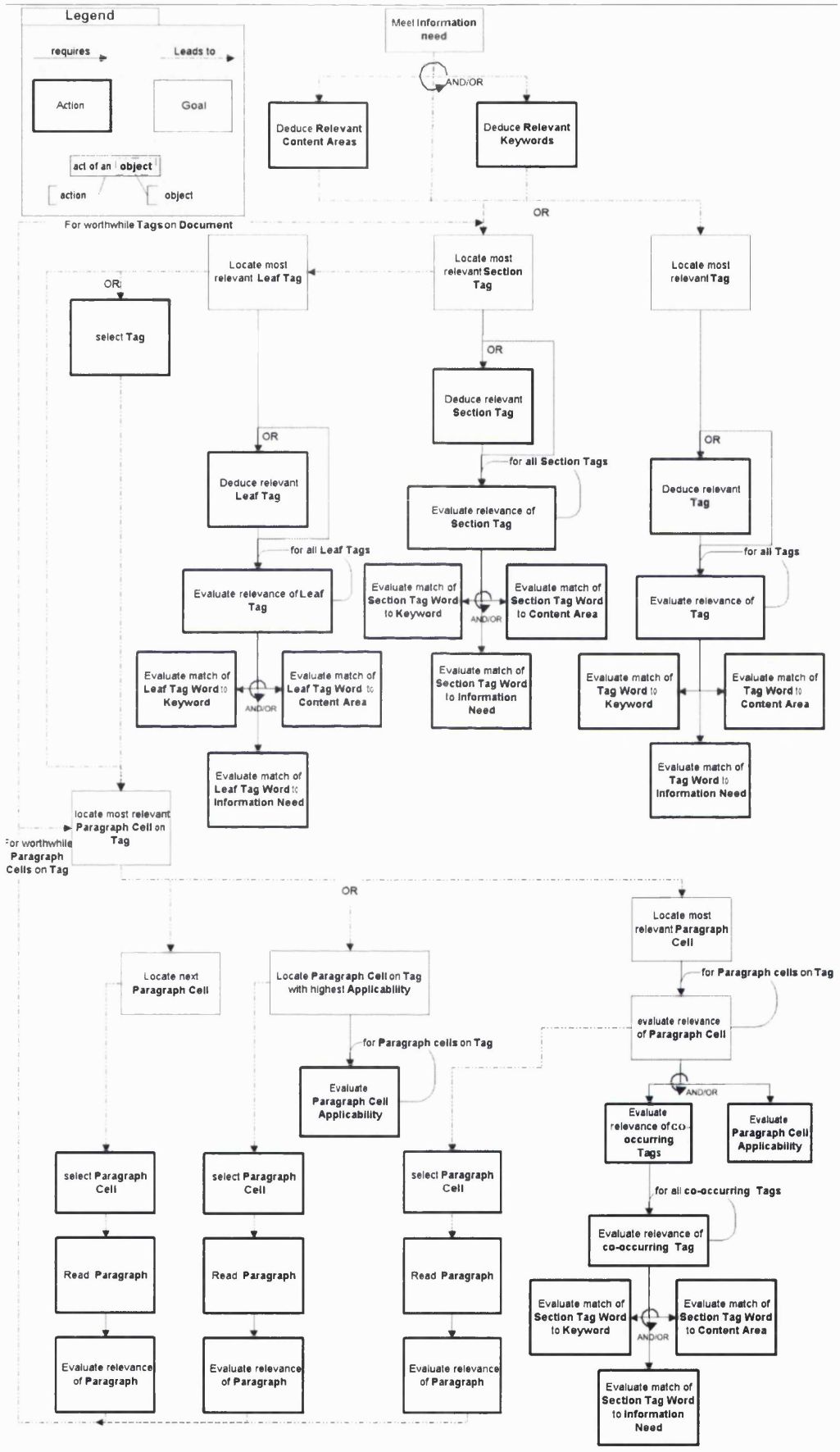


Figure 31 Diagram showing the procedural and goal structure

2.3 Key goals and objects: a simplified task model

The task model laid out in sections 2.1 and 2.2 was simplified to facilitate the making of design decisions. The full task model was examined in detail to identify just those elements vital to efficient goal-directed search. This resulted in a simplified description of the task, shown in Figure 32, involving the key goals and objects involved in goal-directed search using paragraph level metadata. This simplification was necessary because although the detail of the full task model is useful, a simple representation of the task's important elements greatly eases the process of weighing up design alternatives.

Figure 32 is spilt into the two major portions of the task, tag selection and paragraph selection. Tag selection starts with the *information need.* The user can choose a relevant section tag before they choose a relevant tag, or they can ignore the hierarchy and start with choose a relevant tag. Once a *tag* has been selected, a *paragraph cell* must then be selected. There were three different strategies identified by the task analysis, each differing in the information that is used to make the choice. The simplest strategy makes use of no information, and simply picks the *paragraph cell*s in left to right order. The next strategy is to only make use of *applicability*, and the final one uses both the *applicability* and the *co-occurring tags*. Once a cell is chosen, it is read, and then, if the *information need* is not satisfied, the user either goes back to choosing a relevant *paragraph cell* (if there are more to chose) or goes on to find another relevant *tag*.

The model presented in Figure 32 relates closely to the fuller model presented in Figure 31. The 'tag selection' box in Figure 32 is a simplification of the top half of Figure 31. Similarly, the 'paragraph selection' portion relates to the bottom half of Figure 31.

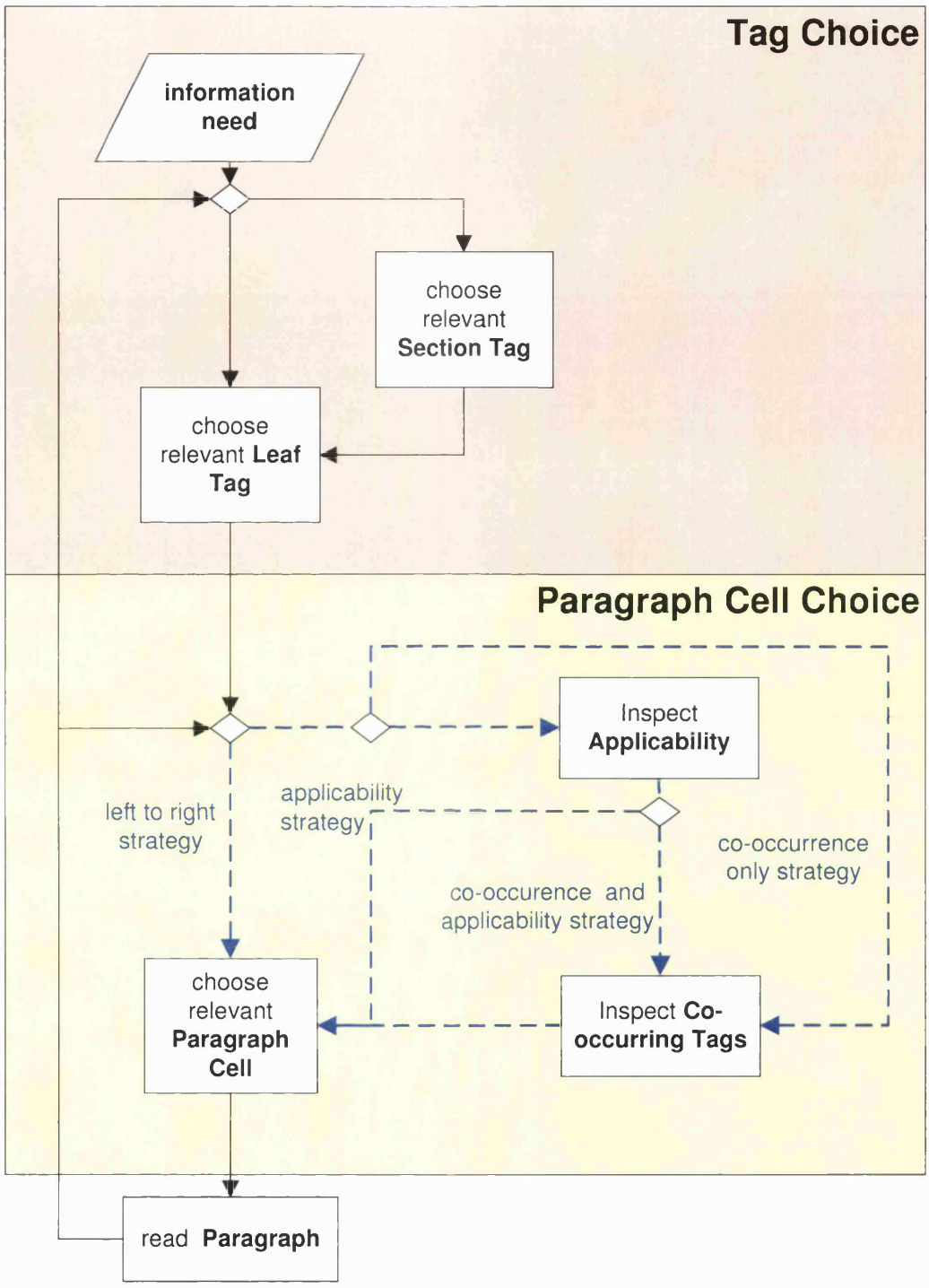


Figure 32 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold),the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines)

2.4 Discussion

This task analysis articulates the details of the use of paragraph level metadata for goal-directed search. These have been aggregated into a simplified model which emphasises the division of the task into tag and paragraph selection. Both tag and paragraph cell selection can be completed using alternative strategies. For tag selection, the strategies essentially differ by whether or not section tags are used to narrow down the search for relevant tags. One would expect the use of section tags to allow relevant tags to be more rapidly identified, hence the design should facilitate the use of section tags. The paragraph cell selection strategies principally differ in how many different types of cue from the visualisation are used in the selection decision. It would seem a reasonable assumption that if more cues are combined into a decision, a better decision will be made. Hence, a design challenge is to allow and encourage their use. This task analysis therefore indicates that design effort should concentrate on making 'section tags' (i.e. the metadata taxonomy), 'applicability', and 'co-occurring tags' visual cues that are easy and efficient to use.

This task analysis and the results from the user trial described in sections 1 and 2 of this chapter, served as a basis for a thorough review of the GridVis design. A number of different designs were generated. These were evaluated and compared by considering their ability to support the key actions and objects, as well and facilitate the use of 'section tags', 'applicability', and 'co-occurring tags'. The design that emerged was a refinement of the same basic design as that used in the user trial (see Figure 34). The design and the rationales behind it are more fully described in chapter 4. A fully functional prototype of this design was constructed and subjected to an empirical evaluation which is the subject of the next section.

3 Experimental evaluation

This section describes an experiment which attempts to provide a full empirical assessment of GridVis's utility for goal-directed search. This small controlled experiment looked at the effect of a fully functional interactive prototype of GridVis on goal-directed search performance and gathered data for use in improving the design.

The experiment required participants to perform goal-directed search with and without GridVis. Twelve participants were each presented with four documents of varying lengths. For each document, each participant was asked to rapidly find the paragraphs containing information relevant to three questions. For two of the documents each participant was asked to use GridVis, for the other two they simply used a scroll bar to navigate through the document. Performance was measured in terms of precision, interactive recall and task time.

3.1 Method

GridVis was tested in the corporate environment for which it was designed; the documents, questions, and participants were all drawn from Sainsbury's. This ensures that the issues and problems uncovered by the experiment are not made irrelevant to the problems of corporations by being rooted in a non-corporate (i.e. academic) environment. For example, it is likely that undergraduate students and corporate knowledge workers differ significantly in their reading strategies. One could assume that knowledge workers are used to reading under time pressure with good knowledge of the background of the documents. On the other hand students taking part in an experiment might be more used to reading with relative little time pressure, having poor background knowledge, and being less motivated to accurately answer queries. Knowledge workers might thus be expected to employ a set of strategies not used by

undergraduates in order to help them read more selectively and save time by, for example, avoiding material that is already familiar.

3.1.1 The sample

The twelve participants in this experiment worked in the Sainsbury's logistics and planning department. This is the department in which, and for which, the documents used in this experiment were written. The participants were familiar with the basic concepts dealt with by the documents and had not previously read any of the documents used in the experiment.

3.1.2 The GridVis prototype and technical set-up

The prototype of GridVis used for this experiment was fully interactive, but only had some of the features described in Chapter 4 section 3.2. The features of the prototype used for this experiment are therefore briefly described here. The visualisation was dynamically generated from XML documents (unlike the static version used in the user trial, described in section 1). Passing the mouse over a column in the grid would highlight all the tags applied to that paragraph, the corresponding paragraph cells and the appropriate paragraph in the iconic document overview (see Figure 33). This allowed the user to inspect the contents of individual paragraphs by placing the mouse over a column and reading off the highlighted tags. A limited form of history information was included in this version of the visualisation; the previously selected paragraph cell was given a red outline. This allowed the users to return to the last point in their investigation. Finally, this version of GridVis included a time counter in the top left hand corner to allow data from log files to be cross referenced with the screen-captured video data.

This prototype did not however have some of the features described in chapter 4 section 3.2.3. The iconic document overview was not dynamically generated for each document, but was a prepared image. The section titles did not appear above the iconic document overview. The history mechanism did not show all of the previously selected paragraph cells. All columns were always shown; there was no dynamic level of detail management feature. Multiple tags could not be selected to show which paragraphs were relevant to them all; there was no co-occurrence automation feature.

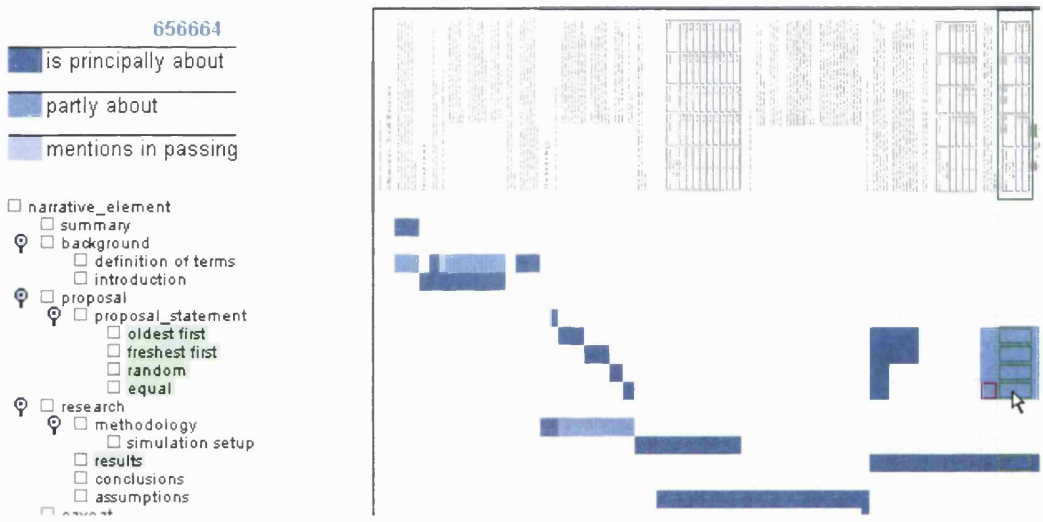


Figure 33 A screenshot of the version of GridVis used in the experimental evaluation

In the experiment, GridVis was presented as a Java Applet in a browser window (see Figure 34). Alongside it are the document window and a frame containing the form elements needed to allow the participant to input their answers and to move on to the next question.

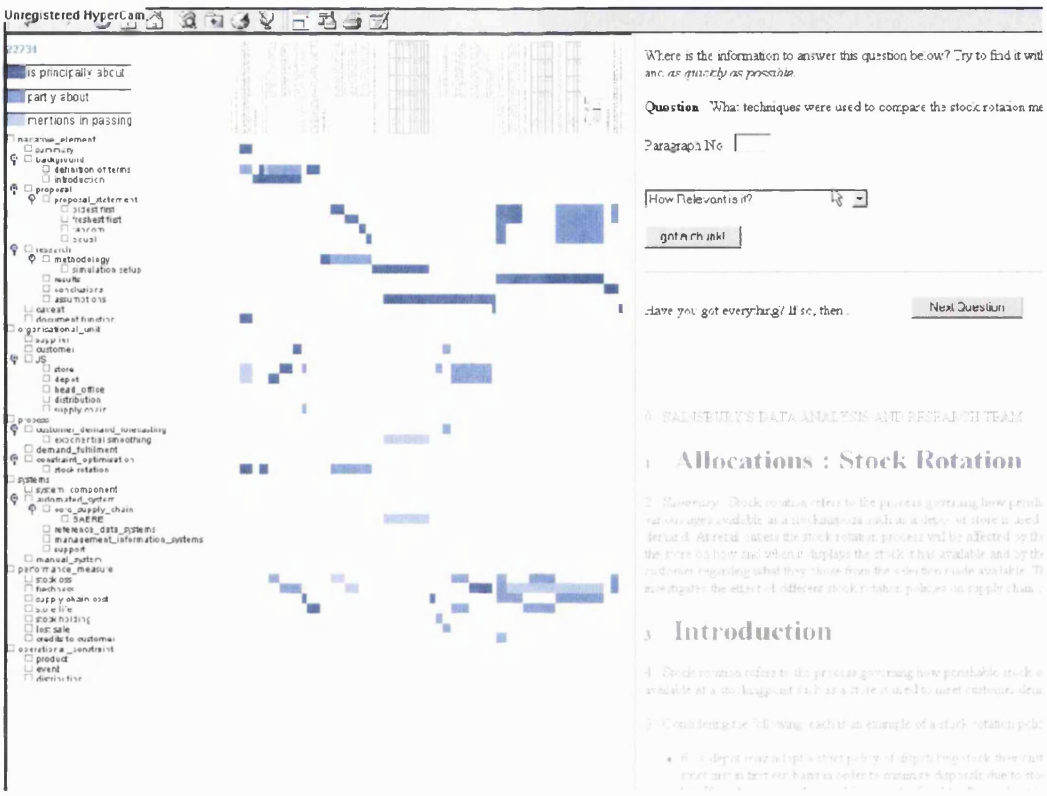


Figure 34 GridVis as used in the experiment.

The scroll bar only condition was presented in much the same way, with a browser window containing the document replacing GridVis (see Figure 35).

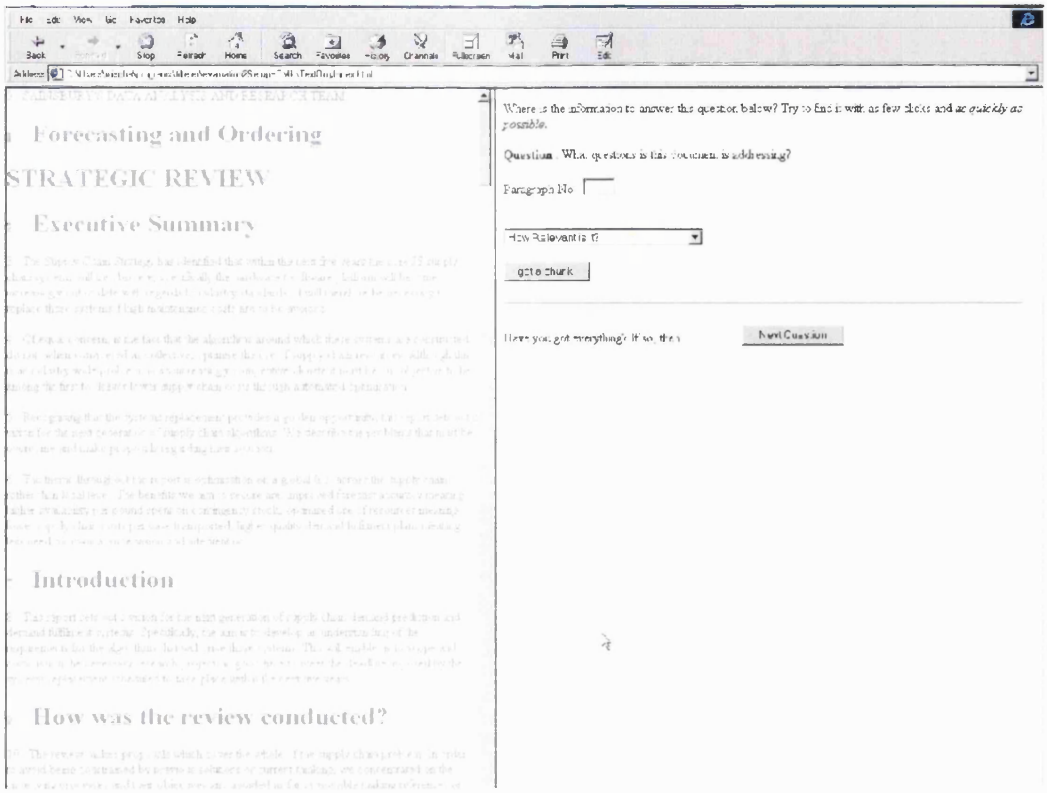


Figure 35 Screenshot of the setup for the 'without GridVis' condition.

3.1.3 The Documents

The documents selected were a set of five reports written by the frontier users; one used for training and four used for the trial. Each document was chosen because it largely conformed to the characterisation, in chapter 3 section 3.2.1, of the kind of document suitable for enrichment with metadata. The documents were also chosen for their varying lengths, three of them being substantially longer (10pgs, 9 pgs,8 pgs) than the other two (5 pgs,4 pgs). Document length was varied to see whether the performance advantage or disadvantage conferred by GridVis would depend on document length. Since GridVis was being designed for use in a corporate environment, the range of document lengths is representative of that observed in the majority of documents used and produced at Sainsbury's (see chapter 3 section 2.2.1). These documents however

had one drawback. One of the requirements for documents suitable for enrichment with metadata, was that documents should be composed of "fairly small spans of text that are self-contained units of information" p. 154. Although these document conformed to this requirements superficially, in that they were written in short terse paragraphs, they also put across substantial arguments which stretched across paragraphs. All five documents were tagged with metadata as described in Chapter 4 section 2.2.

Three questions for each document were elicited from the original audience of the documents. This was done to ensure that the questions used, and their wording, would be as representative as possible of those occurring in real work situations. The author of each document was asked to list the people who read the document. One to two people were interviewed for each document, and were asked to say what they had been looking for, or would look for now, when they read the document. The questions were then examined and excluded if the document did not contain answers to them, or if they required more than the general domain knowledge which could be expected of people working in the frontier users' department.

3.1.4 Relevance ratings

Retrieval performance was measured by comparing the paragraph relevance ratings made by the participants to an aggregation of the ratings made by expert judges. [13]

Before describing the methodology used here to create these benchmark relevance ratings, it is instructive to briefly review existing methods. In evaluations of information

---

[13] Initially performance was scored against the participants' own relevance rating. These were collected at the end of the experiment when the participants had been asked to take time to provide their own relevance ratings for each paragraph against which their experimental performance could be scored. This turned out to be a problematic aspect of experimental design; the time required for this task meant that many of the participants did not take, or have, enough time to do it properly. Once the data had been examined, it was decided that expert judge relevance ratings were necessary if performance was to be reliably measured.

retrieval systems reported in the TREC conferences the author of the topic, which is a description of the information need, makes binary (relevant/not relevant) judgements for a sample of documents (Voorhees & Harman 2000). Voorhees examined the differences in the ratings of multiple judges, and found that although their agreement was low, the impact on the system rankings resulting from these differences was small (Voorhees 1998). However, this finding does not necessarily hold for the evaluation of within-document searches. Additionally, given the small samples in this experimental evaluation, treating variation in valid relevance judgements as error might significantly cloud the results. Kando suggests that multi-level relevance judgements are more natural then the binary ones used in TREC (Kando 2001). The NTCIR (NII-NACSIS Test Collection for IR Systems) information retrieval workshops used a three level relevance scale, and then subsequently a four level one. They also used multiple judges. Two judges rated the documents separately, the primary judge, the topic author, would then make a final decision (Kando 2001).

In the evaluation of GridVis, a procedure similar to that adopted in the NCTIR conferences was used. It was decided to use a four-level relevance scale since this has already been used successfully and it can deliver high definition information about relevance judgements. The relevance of each paragraph was rated on a four point scale, as one of '3 - contains key information', '2 - contains relevant information', and '1 - contains related information' and '0 or unrated – contains no related information'. It was also decided to use three judges for each document. One of the judges was the document's author and the other two were participants in the experiment. However, unlike in the NCTIR procedure, the author of the queries (the researcher) did not decide on the final *single* rating for each paragraph as he lacked domain expertise.

The procedure used to create a single set of relevance judgements used a simple algorithm which was devised from an analysis of the variation between judge ratings. This analysis and the resulting aggregation algorithm are described fully in Appendix 2; only a brief summary will be given here. The analysis found that 36% of paragraphs rated as relevant had no clear relation to the query and that 97% of these paragraphs were only rated as relevant by one judge. An ideal aggregation scheme should omit those ratings which do not relate to the question without excluding those that do. An aggregation algorithm was devised which used objective data (the judge ratings and presence of corroboration) to exclude the majority of paragraphs with little relation to the query, and only exclude a minority of those with some relation to the query.

3.1.5 Experimental design

Twelve participants were asked to perform three goal-directed search tasks on each of four documents, two with GridVis and two without. The documents' presentation order was counter balanced. A within-subjects design was used to avoid the between subject variation that has marred similar experiments in the past (see chapter 2 section 5.1.2.3). A within-documents design would have been desirable for the same reasons but this was not feasible due to possible learning effects that it would introduce.

The goal-directed search task at the heart of the experiment requires the participant to locate, as rapidly as possible, all the paragraphs in a document that are relevant to a given question. They start by reading the document's title and management summary, and then launch into the document using either GridVis or just a browser window. Once the participant locates a paragraph they consider relevant, they rate it for relevance using the four-point scale described in section 3.1.4. They used a web form to enter the paragraph's ID number (which was listed alongside each paragraph) and its relevance rating; unrated paragraphs were assigned a rating of 0. Once the participant felt that they

had enough information to answer a question, or that there were no more relevant paragraphs left, they used a button to move onto the next question.

The experiment began with each participant being given a short list of written instructions describing the experiment and the goal-directed search task. The participant was then shown a training document visualized in GridVis. GridVis's functions were explained by the experimenter. The participant was shown how the metadata taxonomy could be used for locating relevant tags, and how paragraphs' cells could be chosen using co-occurrence and applicability. The participants familiarized themselves with GridVis and the goal-directed search task by using GridVis with the training document to find the paragraphs relevant to three questions.

Once the participant had completed the familiarization run, they moved onto the experimental trials. In these, the participants were asked to repeat the goal-directed search task for three questions on each of the four remaining documents. For two of the documents the participants were asked to use GridVis, for the other two they simply used the scroll bar. When using GridVis, the participants were asked to only make minimal use of the document window's scroll bar and thus rely on GridVis for navigation and search. Order effects were minimized by counterbalancing for the different documents and the with and without GridVis conditions.

Upon completing the experimental condition, each participant was briefly interviewed about GridVis. They were asked about the problems or difficulties they encountered, and questioned on their views of the tool's usefulness.

The participants were then given three paper copies of each of the four documents presented in the experimental trials. They were asked to carefully go through each document entirely for each question, recording the rating of each relevant paragraph in the margin. This was meant to provide a set of ratings against which the paper and

GridVis search performance could be judged. For the sake of brevity, these ratings will in future be referred to as the offline ratings. Several participants did not in fact have enough time to complete this task in the hour they had kept free for the experiment. They were asked to take the documents with them and complete the task in their own time.

3.1.6 Data capture

Data from the experiment was captured using screen recording, interaction log files, and audio recordings. Details of users' interactions were preserved through a continuous screen recording of the entire experimental session, using HyperCam (Hyperionics 2002). Every action (i.e. when the participants clicked on a cell in the grid) for each document and question was automatically logged in a text file. Participants were asked to record the paragraph numbers of germane paragraphs and to rate these paragraphs for relevance. The participant entered this information into an online form (as shown in chapter 4 section 1) which transmitted the information to a Java servlet that recorded the information in a text file. The audio for the entire session was recorded.

3.2 Quantitative results

A quantitative analysis looked for performance differences using the standard information retrieval metrics of recall, precision and task time. A qualitative analysis was then undertaken to examine the strategies used with GridVis and their impact on task performance.

3.2.1 Analysis

The quantitative analysis looked at performance in terms of information retrieval metrics. Interactive recall and precision (defined in chapter 2, section 5.1.2.1) were

initially calculated by comparing the relevance ratings given online (in the 'with GridVis' and 'without GridVis' conditions) to those given offline with paper copies of the documents. Assuming that participants make a perfect selection off-line, paragraphs identified as relevant online, are regarded as false positives if they are not also identified as relevant offline. Thus the number of relevant paragraphs found by a participant is the total number of relevant paragraphs found online minus those not also found offline. This approach had the advantage of embracing individual differences in relevance judgements.

Unfortunately these scores were not felt to be sufficiently reliable because the offline-rating task was not generally done with a high enough degree of care; many of the participants did not take, or have, enough time to do it properly. If the participants had been able to do the offline-rating task thoroughly, then all of the paragraphs found to be relevant online should have been deemed relevant offline. When this was examined it was found that on average 1.2 or 20% of paragraphs deemed relevant online, were not found to be relevant offline. This suggested that offline ratings could not be used to calculate interactive precision and recall. The prompted the use of expert relevance ratings (as described in section 3.1.4).

Differences between conditions were tested for statistical significance using the non-parametric Wilcoxon test (Howell 1997). A parametric test, such as the t-test, could not be used since the data did not appear to be normally distributed when plotted in a histogram (Howell 1997). The Wilcoxon matched-pairs signed rank test is the non-parametric analogue of the t-test for related samples. This is ideal for the data here since document length and use of GridVis were varied within subjects, so all the statistical tests were based on related samples.

3.2.2 Retrieval performance results

The time taken to perform the goal-directed search for a question was on average significantly longer in the 'with GridVis' condition. The mean time taken with GridVis was 162 seconds, while without GridVis the time was 114 seconds. The difference was significant to $p<0.05$, ($Z = -2.197$).

The mean interactive recall calculated using expert judge relevance ratings (see section 3.1.4) showed little difference between the 'with GridVis' and 'without GridVis' conditions; 0.31 for the 'with GridVis' condition and 0.33 for the 'without GridVis' condition. Similarly, interactive precision was 0.53 for the 'with GridVis' condition and 0.57 for the 'without GridVis' condition. The difference between the conditions was insignificant for interactive recall ($p>0.1$, $Z= -.235$) and interactive precision ($p>0.1$, $Z= -.889$).

Document length had a limited effect on performance. The results for this are shown in Table 12. Essentially, document length only significantly affects recall in the 'with GridVis' condition where recall for the short documents is higher than for the long ones.

		Mean for Long documents	Mean for Short documents	P	Z
Interactive Recall	with GridVis	0.09	0.24	P<0.05	Z = -2.136
	without GridVis	0.13	0.21	p>0.1	Z = -1.008
Interactive precision	with GridVis	0.70	0.61	p>0.1	Z= -1.332
	without GridVis	0.58	0.60	p>0.1	Z= -0.178
Task time	with GridVis	157s	183s	p>0.1	Z= -0.178
	without GridVis	117s	143s	p>0.1	Z= -1.125

Table 12 Result for the effect of document length on recall, interactive precision, and task time

3.3 Qualitative results

This section will discuss the results of qualitative analysis of the data collected in this experiment, and a further qualitative investigation. The first section examines the tag selection strategies used, the next looks at the paragraph cell selection strategies, and the final section reports a analytic investigation into the utility of co-occurrence information.

3.3.1 Investigating strategy use

A qualitative analysis of log and video data was undertaken to determine why GridVis did not produce a large performance improvement over the 'without GridVis' condition and to discover what improvements might be possible.

3.3.1.1 Analysis procedure

This analysis started by using the task analysis described in section 1 to look at what strategies the participants might be using to do goal-directed search with GridVis. The task analysis was used to focus the analysis of strategy usage. The task analysis divided

the use of GridVis into two components. Firstly, the user has to find a relevant metadata tag, then they must select from amongst the paragraphs relevant to this tag. The task analysis also suggested the strategies that might be used to make these selections.

This initial set of strategies, taken from the task analysis, was used as a starting point for classifying each tag and paragraph selection action based on the video data and text log files. The classification was refined, using the first few subjects, until it could account for 96% of valid tag selections and 88% of paragraph selections. The final classification of strategies is described in Table 13 and Table 14.

Strategy name	Description of strategy
Keyword mapping	A tag is selected because it matches a keyword from the question
Synonym mapping	A tag is selected because it matches a synonym of a keyword from the question
Related concept keyword mapping	A tag is selected because it deals with concepts related to those in the question
Narrative element mapping	A tag is selected because it refers to part of the document's structure (e.g. 'conclusion') which is likely to contain content related to the question
Unrelated	Not relevant to the question in hand
Unexplained strategy	None of above strategies apply

Table 13 The tag selection strategies

Strategy name	Description of strategy
Left to right	Paragraph cells are selected in left to right order
Applicability only	Paragraph cells are selected in order of their applicability
Co-occurrence only	The relevance of the tags co-occurring on each cell is evaluated. The paragraph cells are selected in order of the relevance of their co-occurring tags.
Applicability and co-occurrence	The relevance of the tags co-occurring on each cell is evaluated. The paragraph cells are selected in order of the relevance of their co-occurring tags. Given cells with similarly relevant co-occurring tags, the cell with the highest applicability is selected.
Unexplained strategy	None of above strategies apply

Table 14 The paragraph cell selection strategies

For each tag selection, if a tag could have been chosen using a strategy then the strategy was given a rating of 1, otherwise it received a rating of 0.

For every paragraph cell selection, evidence from the log files and video data was used to assign scores to the selection strategies. The log files were used to compare the order of cell selections made on a tag, with that which would be predicted given the use of a particular cell selection strategy. For each strategy, a score of 1 was assigned if the order of selections recorded in the log file conformed to the predicted order, and a score of 0 was given if they did not. Separate scores were made using the video data. For example, use of the mouse to highlight and read co-occurring tags, resulted in a score of 1 for the co-occurrence strategies, absence of this resulted in a score of 0. If the mouse was brushed over several candidate cells, this indicated that attention was being given to applicability, so a score of 1 was given to the applicability strategies

This procedure yielded separate sets of scores for the selection order and video evidence; these then had to be merged together. This was done with a simple algorithm that gave priority to the scores derived from the video evidence. For instance, the co-occurrence strategy was still given a score of 1 if a participant was seen to spend time reading the co-occurring tags even if the cells were not selected in the predicted order.

A final transformation was applied to the tag and paragraph cell selection strategy scores. This was necessary to avoid ambiguous selections being over represented. For unambiguous selections a score of 1 is only assigned to one strategy, whereas ambiguous selections result in scores of 1 being assigned to two or more strategies. Hence a transformation was needed which ensured each selection made the same contribution to the overall scores for each strategy. This was achieved by dividing each score by the sum of the scores given to all the alternative strategies for that selection (see Table 15 for an example).

	Left to right strategy	Applicability strategy only	Co-occurrence only strategy	Combined Co-occurrence and Applicability strategy
Raw scores for a selection	*1*	1	**0**	*0*
Transformed scores for a selection	0.5= *1*/(*1*+1)	0.5=1/(*1*+1)	0=**0**/(*1*+1)	0=*0*/(*1*+1)

Table 15 Illustration of the transformation applied to the scores for each selection, to ensure that ambiguous selections were not over-represented

3.3.1.2 Tag cell selection strategies

The overall prevalence of different tag selection strategies is shown by Figure 36. The 'narrative element mapping' and 'keyword mapping' strategies are the most prevalent, with 'related concept keyword mapping' coming next and leaving 'synonym mapping' a poor fourth. Ten percent of the tags selected were deemed to be 'not related' to the question in hand. These are likely to have been due to users selecting cells on rows that did not correspond to the tags they had identified as relevant. A further four percent of tag selections could not be assigned a strategy.

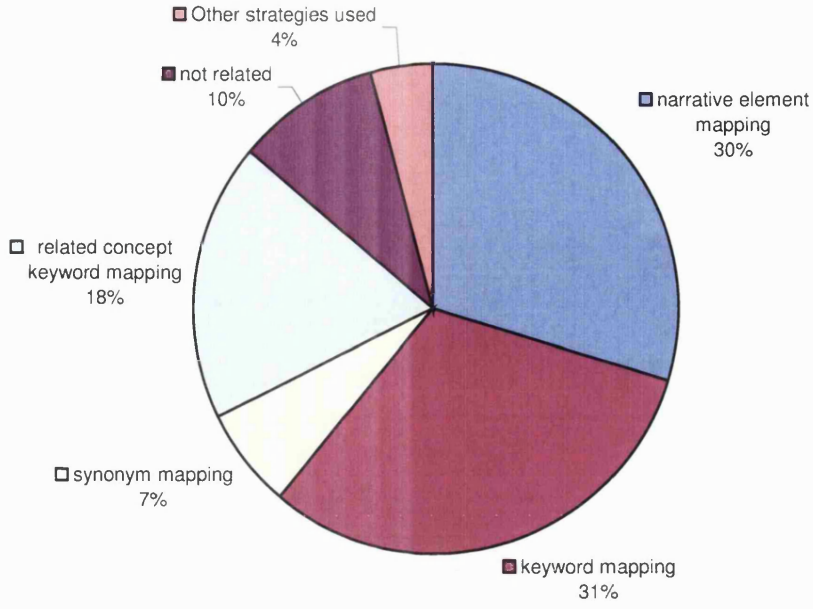


Figure 36 The mean scores for each tag selection strategy across all participants, show as percentage of total score.

The results in Figure 36 suggest that metadata was worded appropriately. Synonyms only have to be used when the questions use a different vocabulary to that used in the metadata. Since the prevalence of the 'synonym mapping' strategy is low and the question wording came from Sainsbury's staff, the metadata must have generally used the same wording as the questions.

There were large individual differences so the pattern of preferences as summarised in Figure 36 does not apply equally to all the participants. An analysis of these differences did not however yield any insights of relevance to this thesis (see Appendix 3).

3.3.1.3 Paragraph cell selection

The overall prevalence of different paragraph cell selection strategies is shown by Figure 37. It shows that the 'left to right' and 'applicability only' strategies are both twice as prevalent as the 'co-occurrence only' strategy, which itself is three times as popular as the 'co-occurrence and applicability strategy'.

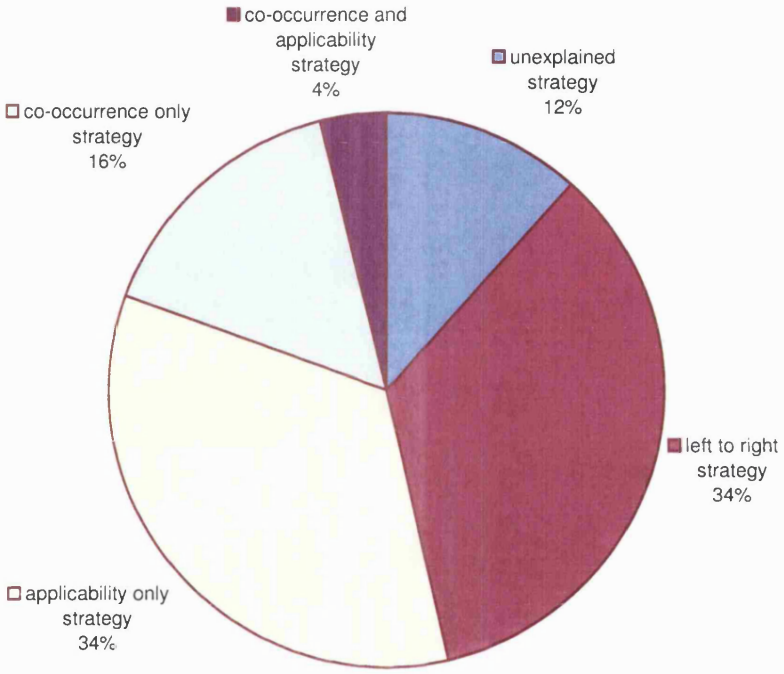


Figure 37 The mean scores for each paragraph cell selection strategy across all participants show as percentage of total score.

As with the tag selection strategies there is much variation across participants. There were two main configurations of strategies; 'no co-occurrence use' and 'some co-occurrence use'. The three participants with a 'no co-occurrence use' configuration (see Figure 38) relied primarily on the 'left to right' strategy while making some use of

applicability. The six participants with the 'some co-occurrence use' configuration (see Figure 39), however, make some use of the 'co-occurrence only' strategy although the 'left to right' strategy is just as popular and 'applicability only' strategy is the one most relied upon.

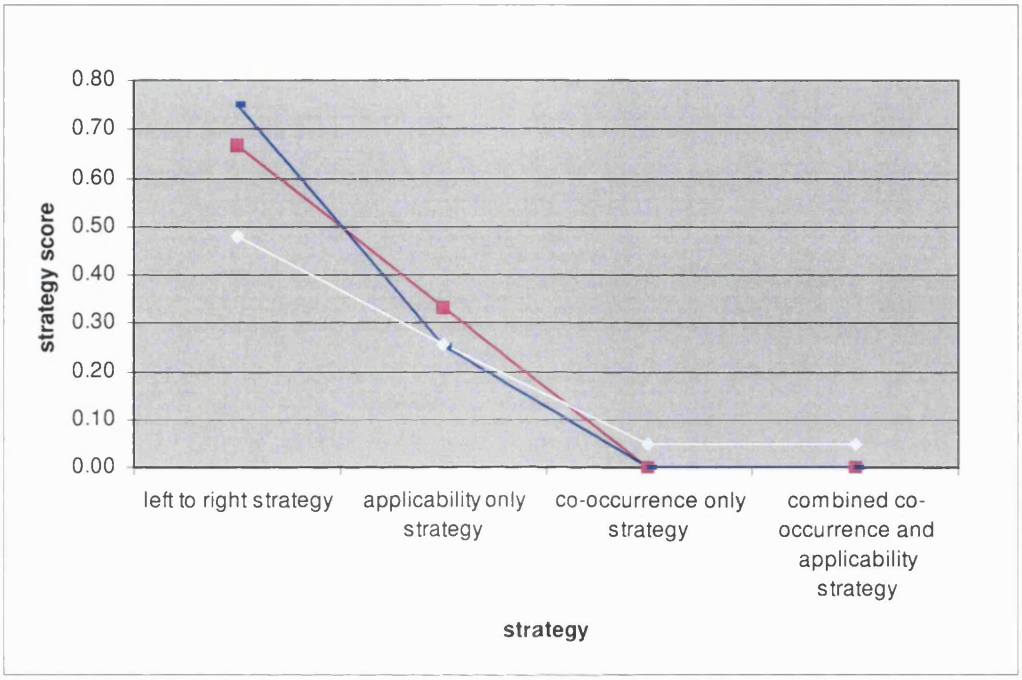


Figure 38 Scores for each paragraph cell selection strategy across three participants with a similar configuration of strategies, shown as percentage of total score.

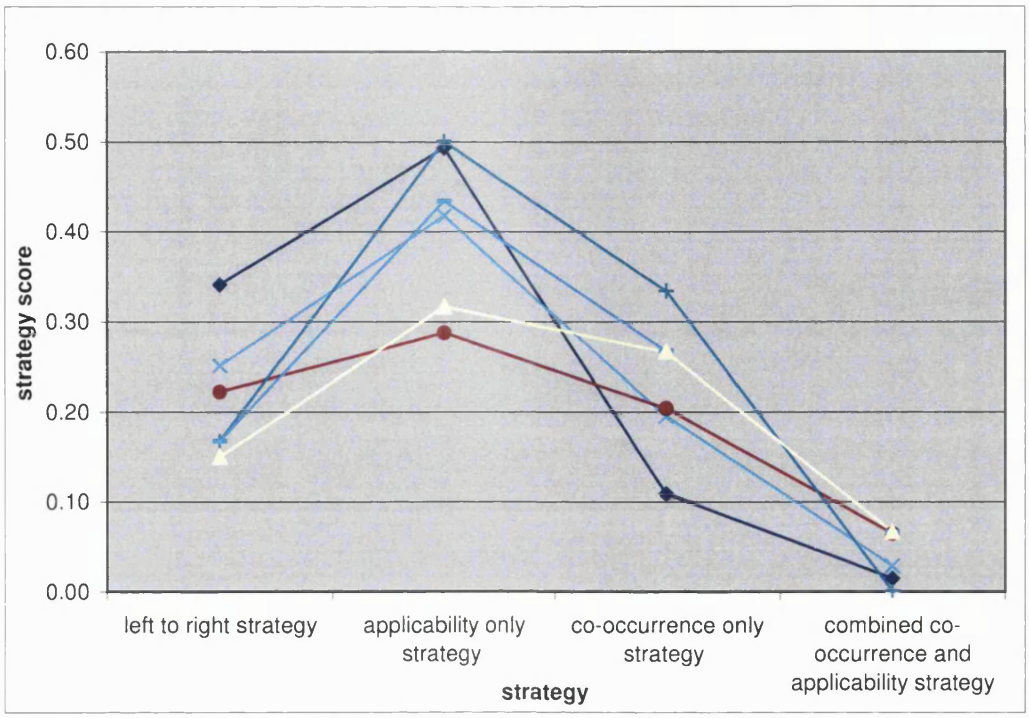


Figure 39 Scores for each paragraph cell selection strategy across six participants with a similar configuration of strategies, shown as percentage of total score.

An interesting aspect of these results concerns what they say about the extent to which applicability and co-occurrence information was used. Overall, applicability was used about 38% of the time, and co-occurrence was used 20% of the time (see Figure 37). Even in the 'some co-occurrence use' group the co-occurrence information was only used in 27% of selections, while applicability was used in 47%. Participants therefore, made relatively little use of co-occurrence information in choosing paragraph cells. Perhaps the poor search performance seen with GridVis is partly due to this tendency not to use applicability and more importantly co-occurrence information.

This hypothesis that the sparse use of co-occurrence explains the low performance was examined by comparing the performance of the 'no co-occurrence use' and 'some co-occurrence use' groups. If co-occurrence information leads to better performance, one would expect the group that uses it to have higher interactive recall and interactive

precision, but lower task time. Contrary to this expectation, although precision and recall were slightly higher, task time was actually higher for the co-occurrence strategy. Nevertheless, none of these differences were statistically significant ($p>0.1$). Either the use of co-occurrence makes no difference to performance or it increases precision and recall at the cost of increasing task time. This suggests that the sparse use of co-occurrence might explain the low recall and precision, but not the poor task time performance.

This ambiguous result, along with the evidence that participants made little use of co-occurrence information prompted further analysis. When the frontier users, and the author of this thesis, used GridVis the co-occurrence appeared, subjectively, to be a powerful way of discriminating relevant from irrelevant material. It was therefore surprising that such little use was made of it, and that it might not result in any performance increase. Further investigation was needed to understand why the co-occurrence information was not being used.

Co-occurrence information might be problematic by being misleading, or by taking too much time to exploit; the data from this experiment only allows the former possibility to be examined. The co-occurrence information could be inherently useless and even misleading by not clearly distinguishing irrelevant paragraph cells from relevant ones. The use of co-occurrence might also be a time consuming distraction from more fruitful activities. Unfortunately, it was not possible to examine the implications of co-occurrence's time cost here, although this question will be dealt with comprehensively in chapters 8 and 9. However, it was possible to examine the effect of the co-occurrence information retrieval effectiveness in terms of traditional precision and interactive recall. This investigation and its results are described in the next section.

3.3.2 Investigating the use of co-occurrence information in cell selection

This section presents an investigation into whether the use of co-occurrence in paragraph cell selection can increase retrieval effectiveness. This was done by exhaustively applying the 'co-occurrence only' and 'co-occurrence and applicability' strategies and comparing the interactive recall and traditional precision achieved with them with that achieved using the 'left to right' and 'applicability only' strategies.

3.3.2.1 Analysis procedure

In this investigation, the tag selection strategies listed in Table 13, p. 220, were systematically applied to each of the twelve questions used in the experiment. For each question every relevant tag was identified. This was done by exhaustively applying each of the four tag selection strategies in order of their overall popularity (see section 3.3), thereby picking out all tags that the strategy dictated were relevant.

For each of these relevant tags, the paragraph cell selection strategies were used to produce a separate ranking of the cells available on the tag. These rankings were based solely on the strategy being used; so cells indistinguishable on the basis of that strategy were given the same ranking (see Table 16 for an illustration). Paragraph cells that had already been selected on a previous tag, were skipped.

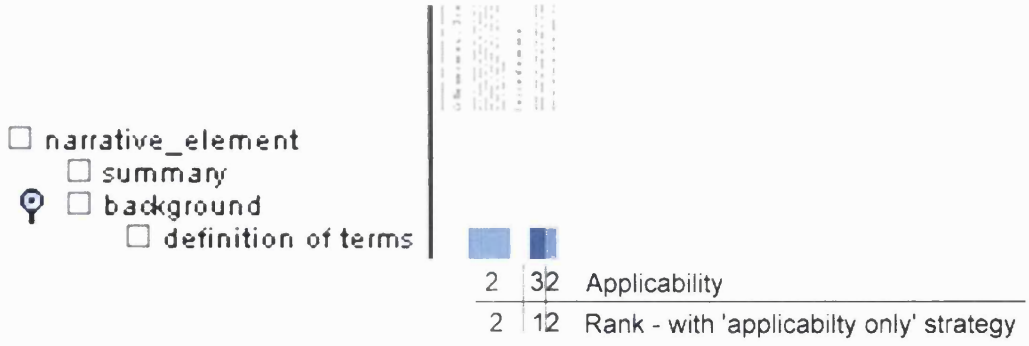


Table 16 An illustration of how the paragraph cells were ranked using the paragraph cell selection strategies.

The left to right strategy was not included as the purpose of this analysis was to examine the utility of co-occurrence information. Therefore it was enough to compare the use of co-occurrence information with that of applicability information.

In order to compare the different cell selection strategies, the top third ranked paragraphs from each group were compared. The ranking of paragraph cells with each cell selection strategy produced three lists of paragraphs and their associated rankings. Since each list contained the same paragraphs, paragraph cell selection strategies could only be differentiated by their rankings of paragraphs. A further complication was that unlike the other strategies, the 'applicability only' strategy resulted in a large majority of paragraphs receiving a ranking of one. The approach taken was to skim off the top ranked third of the paragraph cells. This proportion was chosen because it resulted in a roughly equal number of paragraphs being selected for each strategy. The recall and traditional precision (see chapter 2 section 5.1.2.1 for a definition) achieved with this group of paragraphs for each strategy was then calculated.

Traditional precision can be applied here, if one considers that the tool being used makes paragraphs available via affordances for avoiding irrelevant text. Hence, traditional precision can be calculated by substituting the usual denominator 'Total

number of paragraphs inspected' for 'Total number of paragraphs deemed relevant by the system'.

Traditional precision was calculated here, even though interactive precision was used in the quantitative results presented in section 3.2.2. Interactive precision could not be calculated in this analysis, since the analysis did not produce a human evaluation of a paragraph's relevance. Traditional precision, however has the advantage of being directly comparable with precision measures reported in the information retrieval literature.

3.3.2.2 Results – an exploration of the effect of paragraph cell selection strategy

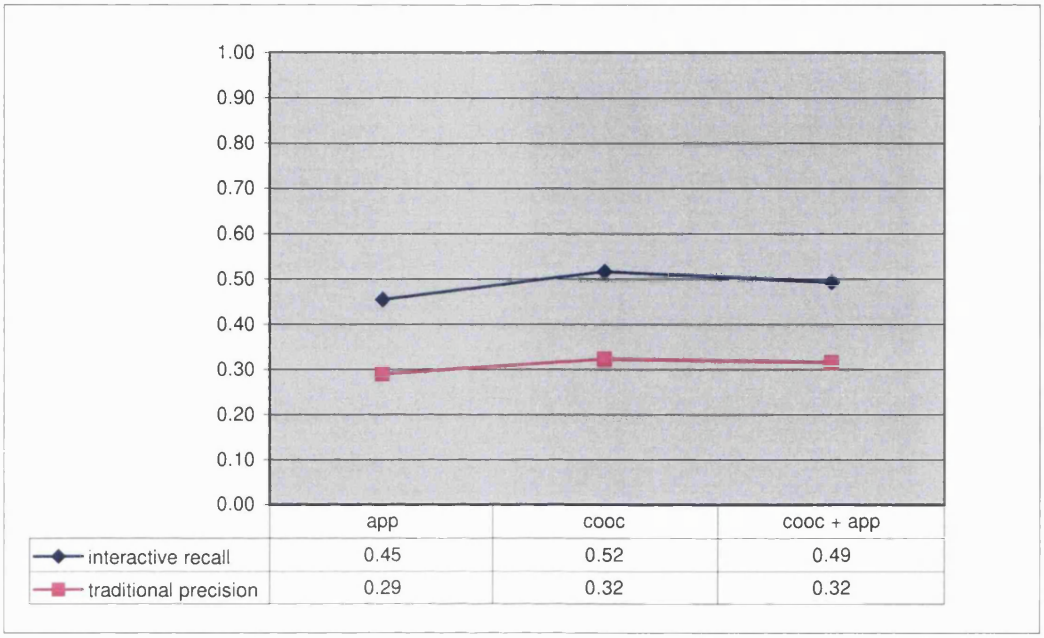
This section reports the results of the exhaustive application of the tag and cell selection strategies uncovered in section 3.3.1. The key question examined is whether the use of co-occurrence information affords benefits in terms of recall and traditional precision.

The use of co-occurrence offers a slight advantage over the use of applicability alone. This advantage is about 10%[14] for both traditional precision and interactive recall (see Figure 40). It suggests that co-occurrence information can indeed offer an improvement in information retrieval performance. Nevertheless, the difference is small. This raises the question of whether it is indeed a noteworthy difference. Unfortunately, since the result for each strategy is taken by pooling the data from all twelve queries and four documents, only one data point is produced per strategy and thus cannot be subjected to

---

[14] For traditional precision, the co-occurrence based strategies have an advantage of $\left(\frac{0.29-0.32}{0.29}-1\right)*100=10\%$ over the applicability strategy. For interactive recall the advantage comes to 17% for the c-occurrence only strategy and 8% for the co-occurrence and applicability strategy.

statistical analysis. Hence, although this analysis suggests that co-occurrence may offer an advantage, the advantage is small and may not be reliable.



	app	cooc	cooc + app
interactive recall	0.45	0.52	0.49
traditional precision	0.29	0.32	0.32

Figure 40 Interactive recall and traditional precision for the top ranked 3rd of paragraph cells using each of 3 different paragraph selection strategies.

Finally, this investigation permitted some light to be shed on the potential of the retrieval performance that a user of GridVis might achieve. If the user were to read all the paragraphs related to the tags selected in this analysis they would achieve a interactive recall of 0.65 and a traditional precision of 0.2. This gives an indication of the maximum interactive recall that could be achieved. The application of the cell selection strategies should then improve the precision while having a minimal impact on interactive recall. If the user were to use the paragraph selection strategies to choose a third of the paragraphs, then, at best, interactive recall drops to 0.52, and precision climbs to 0.32. This performance compares very favourably to the interactive recall of 0.16 found in the empirical experiment's 'without GridVis' condition. Overall, this analysis suggests that if used diligently, retrieval performance with GridVis would be

substantially increased relative to the performance observed by participants in this experiment.

3.4 Discussion

The quantitative analysis suggests that performance with GridVis is broadly similar to that without it. It is only on task time where performance with GridVis is actually significantly worse than when simply using a browser window. Although there were differences in the recall and interactive precision scores between the conditions a statistically significant difference could not be detected. This means that, in information foraging terms (chapter 2 section 6), the rate of gain of information is lower for GridVis.

Figure 41 Metadata tree

The recall scores for both with and without GridVis conditions were surprisingly low. They show that when conducting a rapid goal-directed search of an online document, people locate less than 20% of the relevant information it contains. One explanation for this stems from the type of documents used. They were analytical research reports that dealt with complex ideas and put across substantial arguments which stretched across paragraphs. As such, these documents were difficult to skim, and were perhaps best understood by a prolonged linear reading. Further evaluation should therefore use documents that conform closely to the characterisation of documents suitable for enrichment with metadata, with particular attention to the requirement for 'fairly small

spans of text that are self-contained units of information' (as defined in chapter 3 section 3.3).

Another possible reason for the low recall performance comes down to the inadequacies of screen based interfaces as described in chapter 2 section 4.3.2. To summarise this here, paper has a number of valuable navigational affordances that are missing online (Sellen & Harper 1997). Perhaps, therefore, performance would be substantially increased if the goal-directed search task were to be performed with paper documents. Paper would also be a better benchmark, against which to evaluate GridVis, since paper is still the technology of choice for most reading tasks (as argued in chapter 2 section 4.3.2).

Turning more specifically to GridVis, there are a number of possible explanations for the poor performance observed. Firstly, there is the matter of the users' expertise with GridVis. Although participants claimed to be comfortable with the use of GridVis by the end of the training session, they were still essentially novice users of the interface. It is possible that performance would be much improved by greater familiarity with the structure of the ontology and visualisation techniques used. Secondly, it is not clear to what extent performance was impaired by shortcomings of the prototype's design and implementation. For example, the substantial number of repeat visits to paragraphs could be blamed on the inadequacies of the history visualisation feature. The failure to investigate relevant tags could partly be attributed to the relatively small font size, and the lack of visual features, other than indentation, used in bringing out the structure of the metadata taxonomy (see Figure 41). If GridVis were to be deployed for actual use in a corporate office environment, these kinds of problems could be overcome. The 10% of tag selections which seemed irrelevant could be explained by a difficulty in selecting the correct row in the grid for a particular tag. This problem could be tackled by

encouraging the user to first select the tag of interest so that the corresponding cells are highlighted and by highlighting rows as well as cells. Nevertheless, the case for such possible improvements is difficult to make without evidence of their worth, and it is these problems that could have masked any underlying advantage that GridVis offers.

Another problem concerns the possible distortion of the goal directed search task by the experimental design. The rather small amount of time spent reading, rather than actually digesting the relevant material as they found it, suggests that participants were moving on once they had concluded that something was relevant. If so, the lack of time spent with potentially relevant paragraphs could have contributed to the low precision score.

Another surprising result of this experiment was the higher recall performance with short documents. It was hoped that GridVis would enjoy relative independence from document length. In reality, it seems, document length has a more deleterious effect on performance with GridVis than without it. It is not clear why this should be so, further evaluation work examines this question again in order to replicate the results and better understand its cause.

Qualitative analyses examined the strategies used with GridVis in order to locate design problems that might be limiting performance. This analysis found that there was little use of co-occurrence information. A separate investigation showed that making use of this information could increase recall and traditional precision scores by around 10%. Nevertheless, the data suggest that participants who made some use of co-occurrence might have lower recall and tasks time performance. This lower performance could be a consequence of investing time in the use of co-occurrence information, at the expense of other more useful activities. Further work reported in chapter 6 and 7 will be directed at examining this possibility.

This further work must also look at the utility of applicability information. The focus of the final qualitative analysis was the utility of co-occurrence; applicability was assumed to be of some benefit. The results of the best-case evaluation reported in chapter 7 shows this to be a rash assumption, since it reveals that the 'applicability' strategy shows no advantage over the 'left to right' strategy.

The results of the investigation into the use of co-occurrence information in cell selection, suggest that GridVis has the potential to deliver high retrieval performance. The analysis showed that the diligent application of the tag selection strategies in the use of GridVis could result in much better recall performance than that achieved by participants in the experiment. Nevertheless, this analysis does not take into account the time cost incurred in pursuit of this performance; the high recall performance achieved can only be considered as an upper bound. Further evaluation will more closely examine GridVis's potential, taking both time spent and information gained into account.

4 The problems of empirical evaluation in a corporate environment

The experimental evaluation was unable to reliably detect the effect of GridVis on recall and interactive precision. Given the small magnitude of the difference between the mean scores, and the relatively high variation between participants, a very large sample would be needed to detect a significant difference. Obtaining a suitably large sample size in a corporate environment is extremely difficult for academic researchers.

GridVis was created for use in a corporate environment. Its design has therefore evolved around the corporate material (documents, metadata etc) it uses. These materials, however, mean little to the uninitiated. Also, for the evaluation to relate meaningfully to use in the work situation, representative information needs were elicited from Sainsbury's staff. Only people with the appropriate background knowledge will fully understand these information needs.

Hence, to properly test GridVis, using controlled experimentation, it is necessary to recruit participants from the Sainsbury's staff working in parts of the business for which the documents used are relevant to their work. This means that obtaining more than the twelve willing participants found for this experiment would be extremely challenging. Obtaining the one hundred or more required for a sufficiently powerful statistical test is virtually impossible.

Another problem with using controlled experimentation in a corporate environment is that one can only ask these participants to devote a small portion of their time to an evaluation; an hour at most. This did not seem to be enough time for users to become truly confident with the visualisation tool before doing the experiment. Consequently, this kind of empirical evaluation can only test the performance of novice users.

Although it would be possible to create another set of documents for testing with a more accessible user population (say undergraduates) the results obtained would not necessarily relate back to the corporate environment. As argued in section 3.1 due to differences in motivations and experience, there might well be important differences in the kind of goal-directed search done in a laboratory environment and in a business environment. Since GridVis was created for use in a corporate environment, it is important to verify its effectiveness *in* that environment

5 Conclusions - Implications for the further evaluation of GridVis

A user trial was carried out with an early mock up of GridVis to see how it might support goal-directed search (see section 1). This uncovered a number of design problems and highlighted methodological concerns, which needed to be addressed in the design of any subsequent experimental evaluations. A task analysis of the use of paragraph-level metadata for goal-directed search was conducted (see section 2) to help review the GridVis design. The improved design was fully implemented and subjected to an experimental evaluation (see section 3). This found that performance with GridVis was broadly similar to that with a simple scrolling window, except that task times were significantly slower. These results are essentially inconclusive. The reasons why will now be explored in order to produce requirements for the next round of evaluation.

One of the problems of the user trial and empirical evaluation is that they evaluate a concrete reality rather than any underlying idea. This research aims to evaluate the *general* utility of interactive visualisation of paragraph level metadata for goal-directed search of single documents. The user trial and experimental evaluation looked at the performance of particular implementations with a particular sets of users with particular sets of documents and queries. In an ideal world, this should also be a test of the idea behind the implementation. Unfortunately, it is often possible to argue, as has been the case in the user trial and experimental evaluation, that the reality has not done the idea justice (see chapter 2 section 5.1.2.3). Any future evaluation of GridVis should attempt to avoid this problem by determining the true potential, or best case, of the idea behind GridVis.

The experimental evaluation failed to find a statistically significant effect of using GridVis on recall or precision scores. This did not, however, mean that performance was unaffected by the use of GridVis. There was a large amount of variation, which vastly overwhelmed any persistent differences in performance. As argued above it would not be possible, in a corporate environment, to obtain the large number of participants required to pierce through this fog of variance. This would also be the case for any environments where logistical, or financial, factors make it impractical to recruit suitable participants. The methodology of the next round of evaluation must therefore allow GridVis to be statistically differentiated from the competition using relatively small sample sizes.

The utility of different cell selection strategies was examined in the qualitative analysis. Although the results were suggestive, they ultimately left more questions than answers. The next evaluation needs to be able to assess the utility of trade offs between the alternative cell selection strategies.

The next chapter describes the best-case evaluation methodology. This novel methodology was designed specifically to provide a robust and rich evaluation of GridVis by meeting the requirements layout in this section.

Chapter 6 A best-case analysis evaluation of the potential of within-document metadata for goal-directed search

1 Introduction

The previous chapter described a user study, a task analysis, and an experimental evaluation of GridVis. The quantitative results of the experimental evaluation suggested that performance with GridVis differed from that with a simple scrolling window only in that it was slightly slower when used by novice users. However, qualitative analyses suggested that performance could be improved if strategies were adopted which made more use of the information provided by GridVis. It is argued that the potential utility of different strategies and of GridVis itself cannot be evaluated experimentally in a realistic use setting where there are insufficient participants available for an appropriate experimental evaluation. This chapter defines and describes a new approach, termed a *best-case analysis*, for evaluating the design of an interactive system, which overcomes these problems. It is intended to allow the evaluator to assess, robustly and quantitatively, the potential utility of an application design or class of designs without the need for large numbers of participants. The application of this best-case analysis approach to GridVis is described.

A best-case analysis essentially compares the best-case performance of the design or tool in question with that of a benchmark, which is the most prevalent or effective competing tool. The best-case performance of a tool is the best performance which can reasonably be attained by a human user of the tool. Theoretical techniques are used to

model expert use of the tool and compute the resulting performance. The benchmark performance is either measured empirically using expert users, or computed using modelling techniques. Applied here a best-case analysis compares the goal-directed search performance of human experts using paper to the theoretically predicted best-case performance with GridVis.

The chapter begins with a discussion of the different components of any best-case analysis; this discussion will be followed by an exposition of the approach's scope, advantages and limitations, as well as a brief survey of the different methodologies that could be used in modelling best-case performance. Then a detailed treatment of the approach's application to GridVis will be given. The evaluation will first be outlined. Then a detailed exposition of its components will begin with the materials sections, where the construction of documents and queries used in the evaluation will be described. This is followed by a description of the experiment done to measure human performance with paper.

1.1 Best-case analyses

Here we describe the approach in some detail, examining the alternative methods it could draw upon, and the capabilities and limitations it bestows on the evaluator.

1.1.1 What is a best-case analysis?

A best-case analysis is made up of four components. Firstly, the precise way the tool will be used for the tasks of interest needs to be thoroughly understood. The best-case model needs to be constructed so that the best-case performance can be calculated. The benchmark performance of alternative tools needs to be empirically measured. Finally, the best-case performance must be compared with the benchmark, so that an assessment of the tool's potential can be made.

The first of the four components in a best-case analysis uses a standard task analysis. There are several well-defined types of task analysis (e.g. Carey, Stammers & Astley 1989; Johnson 1992), which produce an explicit and detailed understanding of how a tool would be used. In order to arrive at a valid best-case it is essential to do a task analysis unless the tool and tasks are trivial.

Based on the task analysis the best-case model is then constructed and its input parameters are estimated. The best-case model is a detailed description of an optimal strategy for completing the task. It is constructed from the results of the task analysis using one of several possible approaches (these are discussed in more detail in the next section). The model must be specified down to the level of individual actions whose parameters, (e.g. completion time, probability of success) can be reliably measured and predicted. Once the model has been constructed, reliable values for the input parameters have to be obtained. Some of these parameters can be extracted from the scientific literature, the rest have to be obtained through empirical measurement or simulation. With a fully specified best-case model, and reliable estimates for all its input parameters, the tool's best-case performance can be calculated.

The best-case model alone is interesting, but it can be used to evaluate a technology only when one knows the benchmark performance which it must surpass. This benchmark can be obtained by empirically measuring current expert performance of the task using the current common, or best, alternative tool. Alternatively, the benchmark performance can be determined by a best-case model. The best-case performance of the tool being evaluated can then be compared to the benchmark performance to establish its overall merit. Parameters behind both the best-case and the benchmark can be examined and varied to determine whether there are any particular scenarios in which the new tool is more or less advantageous.

1.1.2 The pros and cons of best-case analyses

The performance of the best-case model is essentially a prediction of the human performance achieved with the tool if it is used in an optimal way. Therefore if the best-case performance of a tool is lower than that of the benchmark, then the tool's utility is very strongly cast into doubt. On the other hand, if the best-case performance is higher than the benchmark one can only draw the weaker conclusion that the tool has the *potential* of out-performing the benchmark. There is no guarantee that users will adopt the optimal strategy; even if they do, the best-case model cannot take into account some factors which might reduce performance such as errors and workplace distractions.

In common with other modelling approaches, best-case analyses allow the evaluator to examine multiple design questions without an escalation in research costs. In the case of substantial design changes, the structure of the best-case model can be altered to reflect design possibilities such as the introduction of functionality to support new strategies for completing portions of the task. The performance differences between the various versions of the best-case model can be used to evaluate the relative merits of these design alternatives.

Another type of design change is one that attempts to improve the efficiency of a particular operation. Proposed improvements can be formatively evaluated by simply examining the effect of changing the model parameters relating to that operation. If, for example, the response time for a query is decreased, the model can be used calculate the consequent reductions to overall task performance. This kind of formative evaluation can be used to set targets for design. The model can be iteratively re-run with progressively larger improvements to particular features, such as query response time. Then design targets can be set at the level which delivers best-case performance that safely outdoes the benchmark performance (see chapter 2 section 8.3 for an example of

this kind of formative modelling). When design changes have been implemented the best-case model can be used also for summative evaluation. The implemented change is incorporated into the best-case model, which can then be used to predict the degree of improvement in performance.

An important advantage of the best-case approach is that it allows robust and rich quantitative conclusions to be made without the need for large numbers of test subjects. There are two reasons for this. Firstly, when the best-case is compared with a benchmark, much lower sample sizes are required to achieve statistically significant results. This is because an experimental comparison involves the sample means from two or more populations, but in a best-case analysis one sample mean (i.e. the benchmark) is being compared to a population mean (i.e. the best-case performance). This economy of subjects is especially valuable for research on interactive technologies in a use setting where the resources available are limited (see chapter 5, section 4). The second reason why best-case analyses make arriving at robust and rich quantitative conclusions tractable is that they enable multiple comparisons of different designs (as described above) without the need for multiple controlled experiments.

As with most evaluation approaches, there are certain bounds and limitations. Perhaps most import is that the task to which the technology is being applied needs *measurable* and *quantitative* performance criteria. Such criteria must therefore constitute a reasonable assessment of task performance. A second limitation is that the task in question must be one that is not completely novel, otherwise there will be no benchmark to test against. In practice this limitation is rarely problematic since new technologies offer new ways (e.g. information retrieval) to meet old goals (e.g. find information). A third limitation is that the task must be simple enough to allow a theoretical prediction of the best-case performance. Although there are various approaches to predicting best-

case performance, and some are suitable for more complex tasks than others, it is unlikely that any theoretical approach would be able to adequately grapple with complex systems which involve social or problem solving elements.

These limitations have implications for both the reach and role of best-case analyses. Firstly, they are limited in the applications and tasks they can handle due to their needs for measurable performance and relatively simple tasks. Secondly, since this approach relies on measurable performance indicators, it is important that other important aspects of performance are not neglected. Consequently, best-case analyses should generally be complemented by qualitative approaches to design and evaluation. For example, involving potential users in the design process can improve the fit to user needs (Ehn 1988). Also, user testing allows problematic aspects of the interaction to be pinpointed; interviews with test subjects can uncover or clarify how the current design causes these problems (Rubin 1994).

1.1.3 Approaches to constructing best-case analyses

A vital preliminary step is to establish what performance criteria are to be used. It is vital that these criteria represent the aspects of the overall task performance that determine the tool's utility. Depending on the type of technology being evaluated, there may be existing standard measures of performance that, for all their faults, can usefully be applied; here the standard information retrieval (IR) performance measures of task time, precision and recall are used. Another possibility is to derive critical performance measures from task analysis, or from a wider analysis which examines the relationships between the tool, its users and other systems (Newman 1997). Standard measures may suffer from not being ideally suited to the task being examined, but they have the advantage of allowing comparisons with other work in the field.

The prediction of best-case performance can be done in a variety of ways depending on the task performance measures being used. Opting for one of the GOMS techniques allows the use of task times and learning time as measures. Different variants of GOMS have been used successfully for making absolute predictions of expert task times (John & Kieras 1996). NGOMSL (Kieras 1996) (see chapter 2, section 5.2.2), for example, has been used to make relative predictions of task learning time.

Another approach is to focus more directly on determining what set of actions constitute an optimal performance of the task. This is needed for interfaces and tasks where good performance depends on selecting the best path through the landscape of interaction possibilities. One way of doing this is through ecological optimisation models which reflect the limitations on the information and resources available to users. This approach to modelling the use of information seeking tools was found to produce good predictions of user performance (as described in chapter 2, section 6.4) (Pirolli & Card 1999). As described in chapter 2, section 8.3, it has been used by Pirolli to compare proposed design changes to a complex information retrieval tool by computing their best-case performances (Pirolli 1998).

1.2 Applying the best case evaluation approach to GridVis

Here the application of best-case analysis to GridVis is outlined and introduced. Hypotheses about the difference in the cost structures (see chapter 2, section 5.2.1) of goal-directed search using paper and GridVis are discussed and an overview of the evaluation is given.

The advantage that GridVis potentially offers goal-directed readers is the ability to avoid reading irrelevant material. GridVis alters the cost structure of the task, increasing the cost of navigation in the hope that less time will be spent on reading irrelevant

material. In contrast, readers of paper spend relatively time little navigating, i.e. turning pages and reading the table contents, but must read many irrelevant paragraphs. What is important here is how these trade-offs affect the overall performance in terms of both the information gained *and* the time spent collecting it.

There are therefore two key assumptions which motivate the construction of GridVis. Firstly, the time saved by avoiding irrelevant paragraphs should outweigh the extra time invested in navigation. Secondly, for there to be an overall advantage GridVis should either increase, or at least not significantly reduce, the relevance and quantity of material found. These two assumptions are examined and tested by the best-case evaluation.

Essentially the evaluation compares the performance of human experts locating answers to questions in paper documents to the task performance given optimal, or best-case, use of GridVis as predicted using a simulation. The next section, section 2, covers the best-case model of GridVis use. The experiment which measures the performance on paper, i.e. the benchmark, is described in section 3. In addition to measuring the benchmark performance, this experiment was also used to gather data needed to estimate parameters for the best-case model.

Both the simulation and the experiment assessed performance on the same set of documents and queries. The documents were industry news digests of the type circulated round a commercial organisation on a regular basis. Three documents, each of a different length (3, 9, 15 pages respectively), were used. The length of the documents was varied to examine the hypothesis that GridVis would perform better, relative to paper, with longer documents. Two queries were devised for each document. In order to promote their ecological validity, these queries were taken from interviews with six Sainsbury's staff. The query difficulty was varied; for hard queries the headings

of relevant paragraphs included one of the concepts used in the query, whereas for easy queries two were included. The construction of these documents and queries is fully reported in section 3.2.

2 Constructing the best case model

This section describes the construction of a computational model which predicts the theoretical ceiling on performance for an expert user of GridVis. The model will be used to complete a goal-directed search task for two queries on each of three documents, and thus enable GridVis' best-case performance to be measured. The documents and queries used are described in section 3.2. The performance metrics used are described in section 4.

The model follows the tradition of 'engineering models' in HCI. The aim of such models is not to produce perfect replicas of human computer interactions, but to produce predictions which are *accurate enough* to answer the questions at hand (Card, Moran & Newell 1983). Since the model is intended to calculate a ceiling on feasible performance its construction takes the line of ecological optimisation analyses which study the optimal case under the assumption that "when feasible, natural information systems evolve towards stable states that maximize the gains of valuable information per unit cost" (Pirolli & Card 1999 p. 3).

The best-case model was built using GOMS and information foraging theory (IFT); these foundations were chosen because used together they can predict information seeking choices and task time. GOMS is a well understood and documented methodology for modelling tasks and obtaining accurate predictions of task time (Kieras & John 1996). GOMS is therefore useful here since time is an important part of the performance metrics being used to evaluate GridVis (see section 4). IFT was needed to model the two major decision points involved in using GridVis. These decision points were revealed by the task analysis in chapter 2 section 2, as shown in the task model in Figure 42, p. 251. Both points require choices to be made between information items (namely tags and cells). Such choices correspond directly to diet and patch selection in

IFT (as described in chapter 2 section 6). Thus IFT can be used as the control structure of the GOMS model in place of selection rules since it is able to predict tag and cell choices. IFT is also suitable here because it deals in reasonable optima; it predicts the optimal set of decisions given the user's limited abilities and knowledge of the task environment.

The first stage in constructing a best-case model is to determine the best strategy for accomplishing the task using the application. The following section (2.1) will describe how such a strategy was uncovered and defined. Section 2.2 describes this strategy at a performance level model which was then implemented to give quantitative predictions of best-case performance.

2.1 Strategy level model definition

From the GridVis task model (see Figure 42), one can see that there are two points at which the user must make a decision. They must choose a relevant tag, and then they must choose from the paragraph's cells on that tag. Any good strategy for using GridVis for goal-directed search must centre on these decisions. In fact, it is practical to split the task into two stages, tag choice[15] and paragraph cell choice. The way these tasks fit together in the best-case model is illustrated in Figure 43.

[15] The word 'tags' is used as a generic term for both document section tags and leaf tags as illustrated in Figure 43

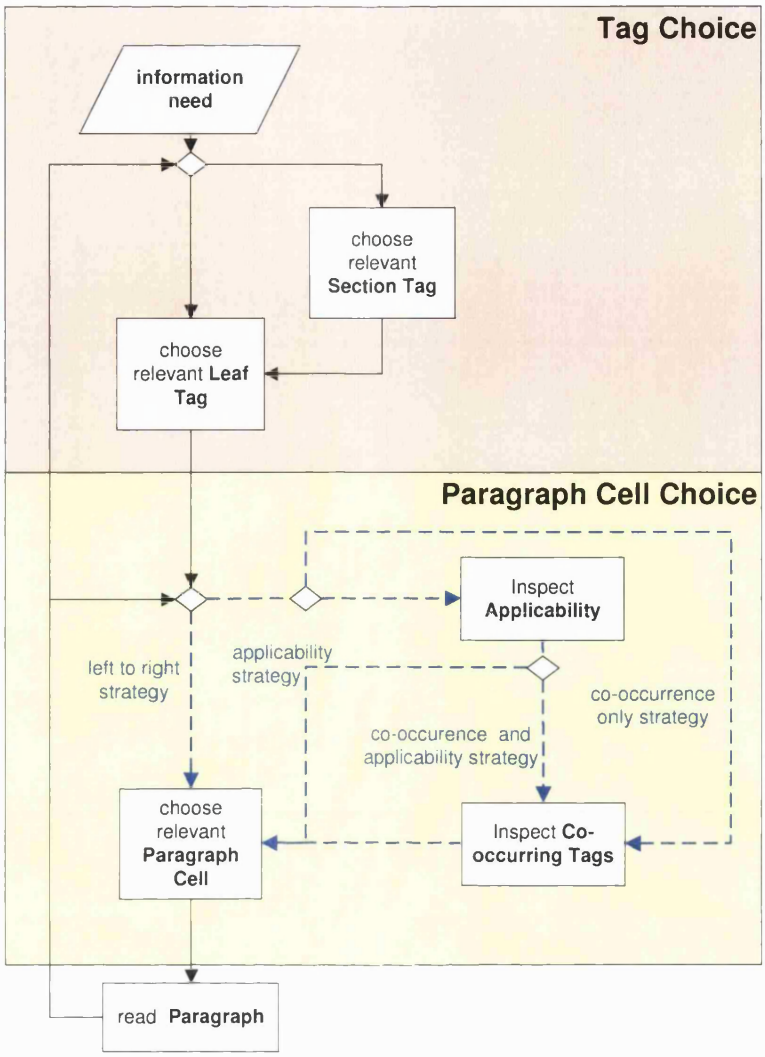


Figure 42 A simplified flow diagram of the task model showing the main portions of the task, the objects (in bold), the input to the task ▱, the decision points ◇, the actions □, and the cell selection strategies (dotted blue lines)

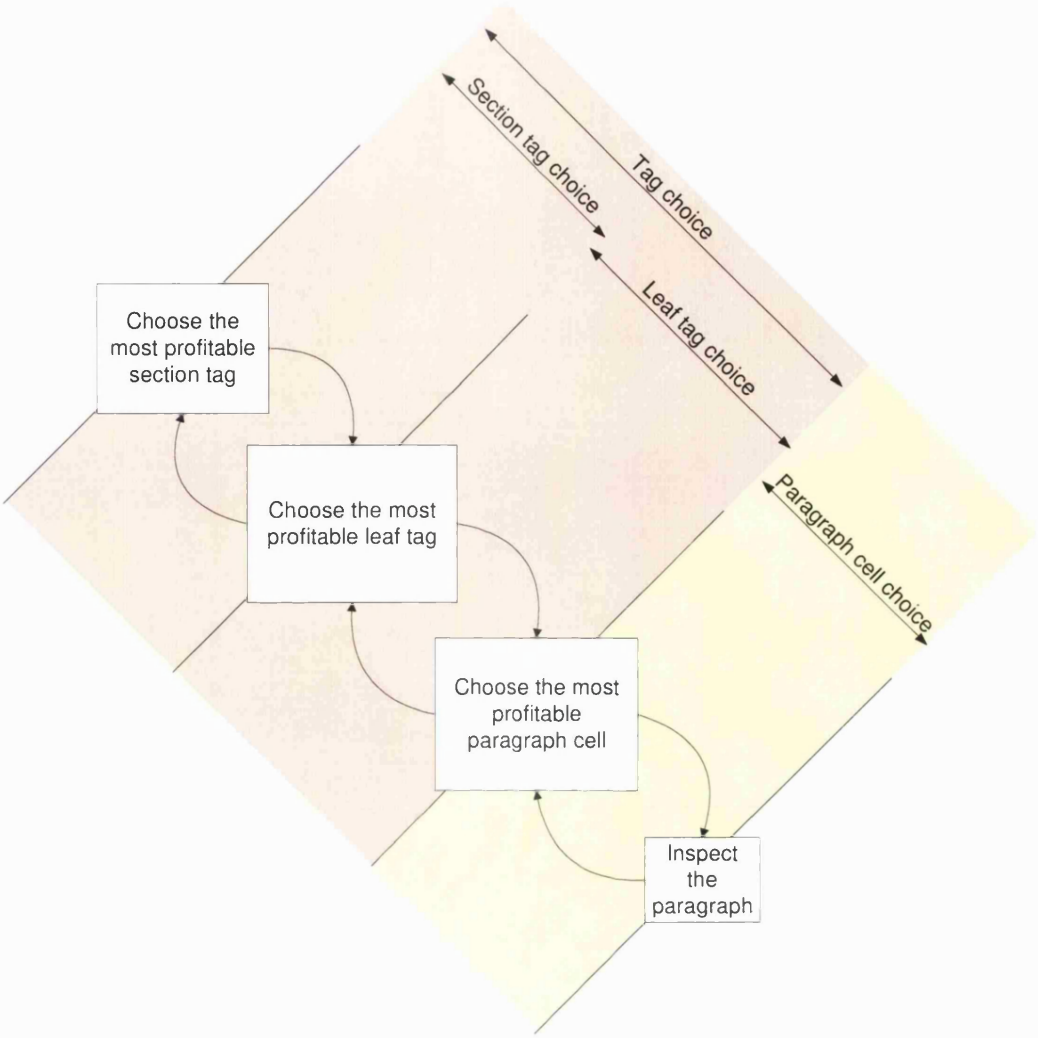


Figure 43 An overview of the best-case model

Since paragraph cell choice is quite clearly constrained by the information available from GridVis, it is simpler and will be dealt with in section 2.1.1. The tag choice strategy will be described in section 2.1.2. First, however, this section will warrant and describe the general strategy adopted in both tag and paragraph cell choice.

The copious research on menu selection can be used to identify two important classes of strategy for choosing tags or paragraph cells. This is because menu selection is analogous to tag and paragraph cell choice in that it centres on the choice of interactive on-screen items. In a review of the menu selection literature Norman describes two

classes of strategy for finding partially specified targets (Norman 1991). In one, called the *optimising* rule, the user evaluates all the alternatives and selects the best, or highest value, option. In the other, called the *satisficing* rule, the user selects the first alternative that exceeds a predetermined threshold value. Users will tend towards the first strategy when the cost of error is great, and will tend towards the latter when time is of the essence.

Here it was decided to use the optimising rule since it would provide an appropriate comparison with performance with paper documents. In the paper experiment subjects tended to inspect the entire document (as reported in chapter 7 section 1.1) despite being given a table of contents, and being asked to work rapidly (as they normally would). Their search was therefore exhaustive, minimising the chance of error, and thus analogous to the optimising rule.

The optimising rule alone is, however, a rather underspecified strategy. At this point information foraging theory becomes of great value. Its 'optimal diet selection strategy' clearly defines a rational search strategy which conforms to the optimisation rule and, to some extent, describes how people actually search for information (Pirolli & Card 1999). Information foraging theory, the optimal diet selection algorithm, and application of these to GridVis will therefore now briefly be explained, although for a more in-depth description see chapter 2, section 6.2.

Information foraging theory is a predictive theory of information seeking. It is premised on the notion that users tend to maximise the amount of valuable information they obtain, while minimising the cost of the interaction required. This commonsensical idea leads to the more concrete assumption that the user is maximising the rate of valuable information gained; in other words they are maximising the profitability of their interactions. This assumption is theoretically fruitful since it allows a choice between

alternative interaction possibilities to be made on the basis of their profitability. This has enabled the construction of quantitative models of information seeking which successfully predict and explain human information seeking as noted in section 6.4 of chapter 2.

As argued in section 6.1 of chapter 2 information foraging can be treated as a repeating cascade of patches and items. The task model for GridVis depicted in Figure 43 can be recast in this mould as in Figure 44. The user begins by selecting a section tag; here the section tags are treated as patches and the leaf tags as items. They then must select from among these leaf tags which are now treated as patches of paragraph cell items. Finally, they must select paragraph cell items and inspect the corresponding paragraphs.

In Figure 44 the user can either make a selection and move down the cascade or go back up the cascade to find a more profitable patch. The optimal diet selection algorithm defines the rational behaviour for making selections and thus for moving down the cascade. It also defines one of the two conditions which trigger a move back up the cascade. Before describing the second condition for going back up the cascade, it will be instructive to describe the optimal diet selection algorithm and the conditions it imposes on movements through the cascade.

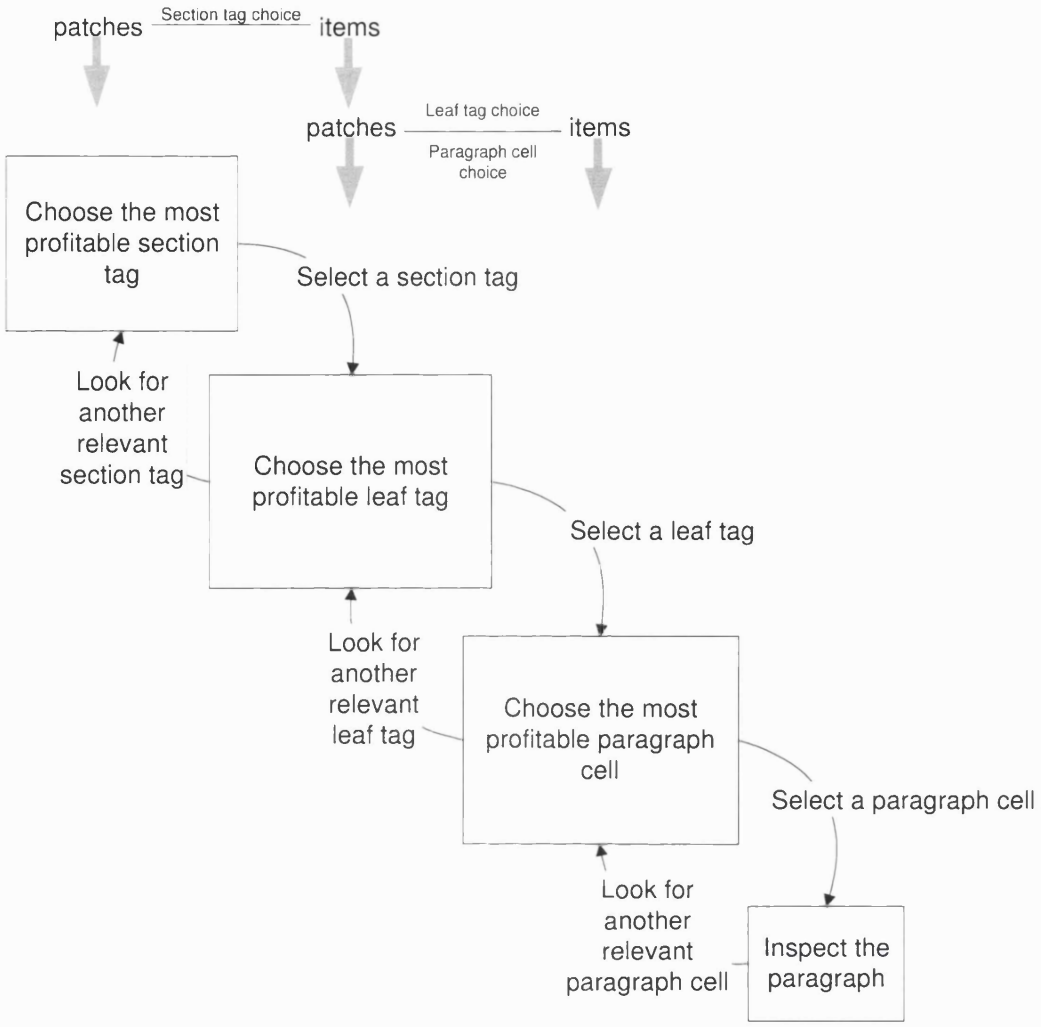


Figure 44 The use of GridVis seen as a cascade of patches and items

The optimal diet selection algorithm, which was derived from foraging theory, computes the optimal, or most profitable, set of information items to select. Essentially, the costs and returns expected from each information item are evaluated and used to calculate profitability. The items are then selected in order of profitability until their profitability sinks below the profitability of the diet so far consumed. Since this algorithm is provably optimal (Stephens & Krebs 1986), it can be used to model the best selection strategy.

Here the diet selection algorithm is described as it applies to GridVis. For GridVis, the optimal diet selection algorithm selects information items, namely tags or paragraphs cells, until the inequality in Equation 3 (derived in section 6.2 of chapter 2) no longer holds true. At this point the forager moves back up the cascade shown in Figure 44 to look for another worthwhile patch.

$$R(k) = \frac{\sum_{i=1}^{k} g_i}{t_b + \sum_{i=1}^{k} t_{wi}} \leq \frac{g_{k+1}}{t_{wk+1}} = \pi_{k+1}$$

Equation 3

The inequality in Equation 3 defines the conditions in which the k+1[th] item can be added to a diet of k items. A given set of information items is abandoned when π_{k+1}, the expected profitability of the next potential selection, is smaller than $R(k)$, the profitability of all the items so far consumed[16]. The information gain expected is, g_i. t_{wi} is the expected time cost of consuming item i, and t_b is the mean time taken to choose the next patch; t_b is assumed to be a constant for each of section tags, leaf tags, and paragraph cells [17]. A similar assumption is made by Pirolli and Card in modelling the use of an information seeking tool (Pirolli & Card 1999).

The optimal diet selection algorithm fully defines how a rational user moves down the cascade, shown in Figure 44, but only gives one of the two conditions which can trigger

---

[16] $R(k)$ should, if the items are consumed each time a choice is made, be calculated using the actual rather then expected profitabilities.

[17] In foraging theory t_b is expressed as an encounter rate, $\lambda_{itemType} = \frac{1}{t_{b_{itemType}}} \cdot \lambda_{itemType}$ varies across alternative items types. This is important because, in biology, encounter rates vary for different types of item (e.g. grass vs. cabbages). However the alternative information items in GridVis are all equally accessible, thus the rate of encountering items, t_b can be assumed to be constant for alternative items within a patch.

a move up the cascade. This is when the inequality in Equation 3 is broken. The second condition is that π_{k+1}, the profitability of the next potential item, is smaller than that of the next patch one step up the cascade. So, for example, a user should abandon a tag's paragraph cells if the most profitable unselected paragraph cell is less profitable than the current section's next most profitable unselected tag. If the forager did not abandon the tag at this point then the overall profitability of their diet would be lower than it would have been if they had moved on to a more profitable tag. This means that each time the user selects a patch, and starts to forage through its items, they have to know the profitability of the next, or runner up, patch, so that they can abandon the items in the current patch that are less profitable than the next patch.

So far, the strategy for choosing information items has been defined. This relies on estimates of the information scent of alternative information items[18]. The strategy level model must therefore also define how the user is to efficiently evaluate information scent (g_i). The following two sections will address this question for both tag and paragraph cell choice.

2.1.1 Paragraph cell scent evaluation strategy

Paragraph cell scent needs to be evaluated when deciding which cells to choose. There are three pieces of information that usefully might be included in paragraph cell scent:

1. Applicability of the cell being considered for selection,
2. Relevance of other tags in the column (paragraph) of interest (i.e. co-occuring tags),

3. The applicability of the co-occurring tags to the paragraph being considered.

The locations of these pieces of information are shown in Figure 45.

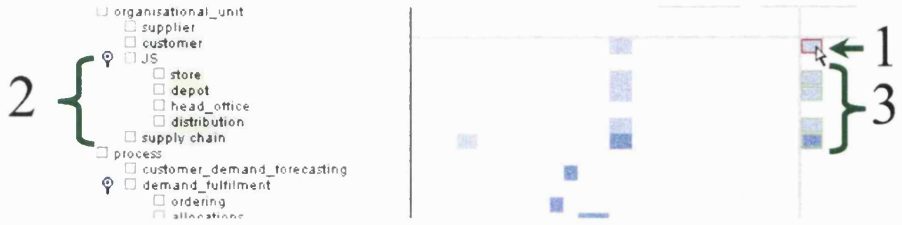


Figure 45 Illustration of the three different types of information available for judging the information scent of paragraph cells

Given the simple nature of the task, the different possible strategies were worked out analytically. By considering the ways in which these three pieces of information *can* be used, the five theoretically possible strategies for paragraph cell choice were identified (see Table 17).

Strategy	Information used
I	None
II	Applicability of the cell being considered for selection.
III	Relevance of other tags in column (paragraph) of interest
IV	Applicability of the cell being considered for selection. Relevance of other tags in columns (paragraphs) of interest
V	Applicability of the cell being considered for selection. Relevance of other tags in columns (paragraphs) of interest The applicability of the co-occurring tags to the paragraph being considered.

Table 17 The different possible strategies for cell choice

The fifth strategy puts a very large demand on the user's working memory. As a result the author found it impossible to follow this strategy when using GridVis. It is thus not

---

[18] Here, and in the rest of this chapter, information scent is defined as the perception of the amount of valuable information to be gained from a particular interaction path. This definition is narrower than that given in chapter 2 section 6 where information scent was defined as the perception of both expected,

a strategy that even an expert user can be expected to follow and is therefore not considered further in this work. The relative advantages of the remaining four strategy will be examined below in chapter 7 by running the best-case model with each of them separately.

2.1.2 Tag scent evaluation strategy

In order to decide between tags, the user must first evaluate their scent. However a simple GOMS analysis reveals this to be a highly costly process (see Table 18). Given the 100 or so tags in a single document, reading every tag would take over two minutes. The strategy must therefore minimise the number of tags read by only attending to those likely to be relevant.

Operator	Time	Source for operator times
Get-from-task <tag name>	340ms	Time to read a word (John & Newell 1989)
Compare to Query	1200ms	Mental operator (Card, Moran & Newell 1983)
	Total 1440ms	

Table 18 GOMS analysis showing the cost of reading tags

An informal study was used to identify a tag scent evaluation strategy that was both realistic and efficient[19].

This informal study revealed a number of strategies. In general, the users would start by trying to remember a relevant region in the metadata tree. So for example, a user may

---

value and cost. Here, for clarity, it has been necessary to separate the discussion of the estimation of expected value, and cost – thus scent is only used to refer to the perception of expected value.

[19]Two frontier users were asked to repeatedly undertake the selection task using an instrumented version of GridVis. They were given six questions. After each tag they selected, they were asked to comment on how they had chosen the tag, and once they had finished the task their strategies were the subject of a further discussion.

GridVis was instrumented to time and record the leaf tag selections. The users were presented with a query. Once they had digested the query, they clicked on a 'start' button to reveal the tags and start a timer. Once they located a relevant tag, they clicked on it. This selection caused the tags to be hidden again and the selection time to be recorded. The user then reselected the 'start' button and chose another relevant tag and so on until they felt they had selected all the relevant tags. They then clicked the 'next question' button.

want to find tags for competing supermarket chains, so they look roughly three quarters of the way down the list, where they remember these tags to be. They scan that region for relevant tags. When they have finished with a section they try to think of other regions that might be relevant and inspect them. Once they have exhausted this tactic, they will rapidly scan the entire list several times looking for relevant tags that 'pop out'. Also, as a last resort, one of the frontier users decided to methodically follow the hierarchical arrangement of tags in order to locate relevant parts of the taxonomy[20].

These results suggested a plausible best-case strategy. The first strategy attempted by both participants was go directly to a potentially relevant region. They only moved on to other strategies because their lack of familiarity with the taxonomy meant they weren't sure what the other parts of the taxonomy contained. So given knowledge of the taxonomy, the region strategy was the frontier user's favourite. In terms of GOMS task completion times this strategy is far superior to the more traditional tree traversing strategies. Since one can assume that a truly expert user would be very familiar with the taxonomy, this strategy can be used alone as the best-case strategy.

2.1.3 Assumptions used in strategy level model

Modelling human activity always carries with it a trade-off between realism and tractability. In defining a concrete and modelable expert user strategy a number of assumptions have been made. These assumptions concern both how the document text is read and how GridVis is used.

---

[20]Another interesting aspect of these results is that they contradict the intuition that the taxonomy's hierarchical structure would be useful to users. It seems that the sequential grouping function of the window explorer style hierarchy display is in fact the most important affordance of the metadata tree. Accordingly, perhaps the visualisation should more directly support this by, for example, using different background colours in different sections.

One feature of human interaction which is hard to model is the dynamic switching between strategies that was observed in the paper reading experiments. There are three points at which the user might want to dynamically switch strategies. Perhaps the most important example is that the user might sometimes switch from GridVis to other more conventional navigational aids (e.g. the scroll bar) when these are likely to be less costly. A second involves the choice between paragraph cell choice strategies; the user could opt for cheaper ones when they are more appropriate (such as when there are few paragraph cells on a tag).

The assumption that this strategy switching does not occur has the advantage of simplifying the model, reducing the number of ad-hoc parameters. It also allows the relative merits of some of the available strategies (such as the alternative paragraph cell choice strategies) to be evaluated; this then allows for a principled consideration of the opportunities for strategy switching. The disadvantage is that it could be argued that performance would be improved if strategy switching was modelled.

The best-case model attempts to capture the essence of the affordances offered by the interactive visualisation of paragraph-level metadata. GridVis includes peripheral features, such as the ability to select paragraphs from the iconic document overview, the display of article titles, and the dynamic level-of-detail feature. These are not included in the model because they would add complexity and are not integral to what is actually being evaluated, which is the interactive visualisation of paragraph-level metadata.

The current implementation of GridVis also has certain shortcomings which could be overcome and should therefore not be modelled. One major shortcoming is that for both the long and medium length documents, the list of tags is too long to fit on one screen. This problem could be overcome in various ways that only minimally alter GridVis's cost structure. One possibility would be to use a fisheye display which only keeps the

region of interest at a readable size (see (Bederson 2000) for an implementation). Another shortcoming concerns the current history visualisation, which does not clearly indicate which paragraphs and tags have already been visited. It became clear that this was particularly important during the construction of the GOMS models, since without it much time is wasted. Including these shortcomings in the model would, as it were, be selling GridVis short.

2.1.4 Summary of the strategy level model

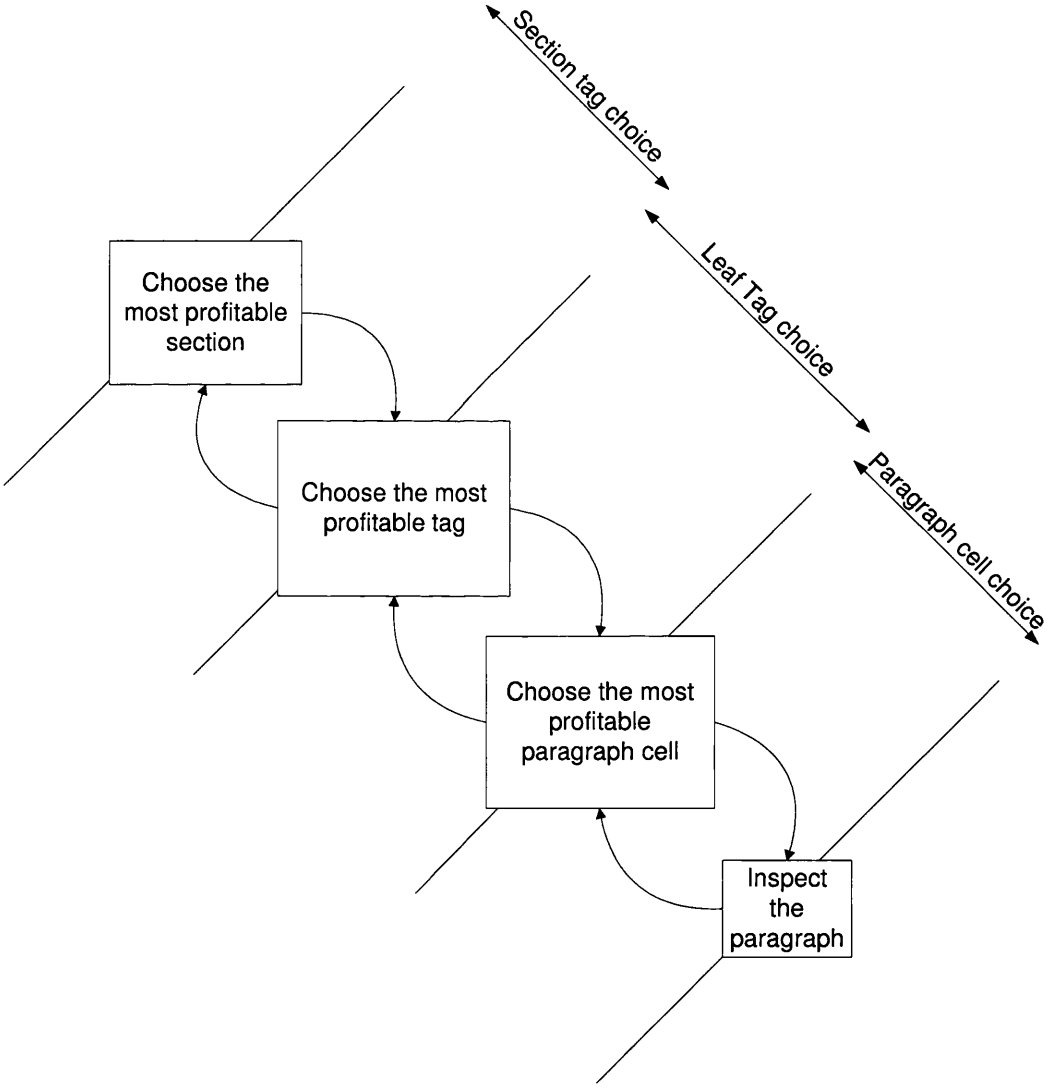


Figure 46 Overview of GridVis best-case model

An expert user of GridVis would begin by selecting a relevant tag. This involves using long-term memory to decide on the section of the taxonomy most relevant to the query. The leaf tags within the section are read to ascertain which is the most relevant. A tag is then chosen and selected, highlighting the related paragraph cells.

One of four different strategies can be used to select the paragraph cells relevant to a particular tag. These strategies vary in the number of cues they use as information scent. The model can use each of these strategies separately in order to allow their relative advantages to be ascertained.

Once a paragraph cell is chosen, it is selected, the resulting displayed paragraph is scanned and if relevant, read in some detail. The best-case model then returns to GridVis, and re-inspects the paragraph cells relating to the previously selected tag. If the cells are worthwhile, i.e. Equation 3 holds, these are selected and their corresponding paragraphs are read. Otherwise, the user re-inspects the section from which the last tag was selected. If there are any other worthwhile tags, these are exploited, otherwise the user turns to the next most profitable section. This process continues until there are no more worthwhile sections.

2.2 Performance level model definition

Having defined the best-case model at the strategy level, the operational details must now be described. However, before diving into the details of the performance level model its structure will be outlined.

The model, which was implemented in Excel using Visual Basic, follows the strategy described in section 2.1. Its structure, illustrated in Figure 46, can be broken down into the choosing of: section tags, leaf tags and paragraph cells. Since the choosing section tags is similar to choosing leaf tags, the model is the same for each; the high-level

structure of the model will thus be discussed first in terms of how tags are chosen and then in terms of how paragraphs cells are chosen. However, both tags and paragraph cells are chosen using the diet selection algorithm which requires data on expected gain, i.e. information scent, and expected cost of pursuing a tag or paragraph cell, i.e. exploitation cost. Thus, the performance level model must compute information scent and expected exploitation costs for both tags and paragraph cells.

Tag information scent was calculated by assuming *perfect* information scent. Perfect information scent is simply the sum of the relevance values of every paragraph that can be reached from a particular tag or cell. Hence the model has an oracle which indicates precisely how much information can be found using each tag. The use of perfect scent for tags and arguments for its appropriateness is discussed in section 2.2.2.1.

A tag's exploitation cost is calculated in two parts, the estimated choice time, and the estimating reading time. The calculation of estimated choice time is described in section 2.2.3.2. The estimated reading time is derived from data on typical paragraph reading times extracted from data provided by the paper performance experiment (see section 3). The actual choice time is calculated using a GOMS model of tag choice described in section 2.2.3.1.

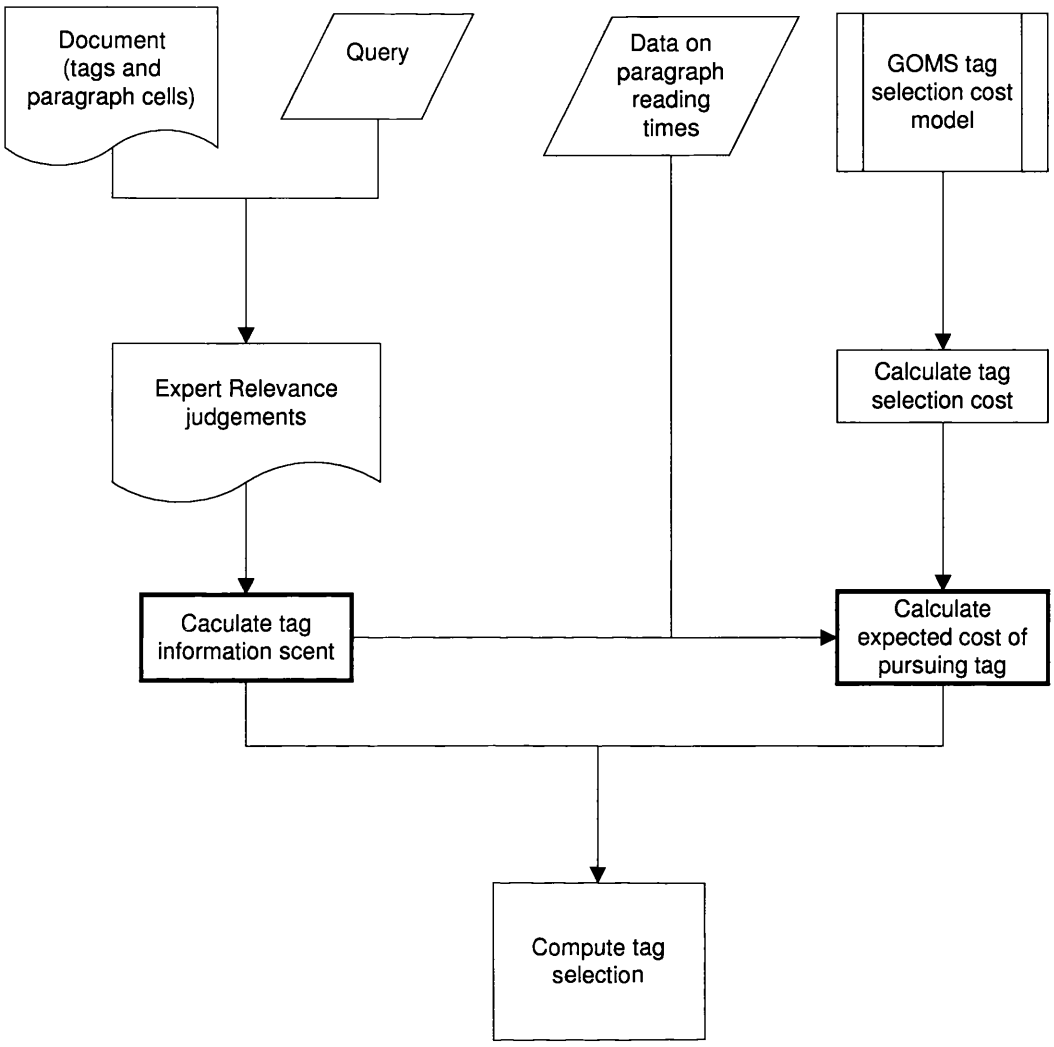


Figure 47 Overview of the tag choice portion of the performance level model

The costs of paragraph cell choice are computed in a very similar manner, but information scent is calculated very differently (see Figure 48). Instead of assuming perfect scent, the model *simulates* the calculation of information scent. Simulating the scent gathered by each of these strategies allows their effectiveness to be examined both in terms of their cost and their retrieval accuracy.

Simulated information scent is based on the information displayed by GridVis. A paragraph cell's information scent is calculated with reference to its applicability and the information scent of related tags (see section 2.2.2.2). The expected exploitation cost

of a cell is derived from the data on typical paragraph reading times (see section 2.2.3.2). The actual cost is calculated using a GOMS model of paragraph cell choice (see section 2.2.3.1).

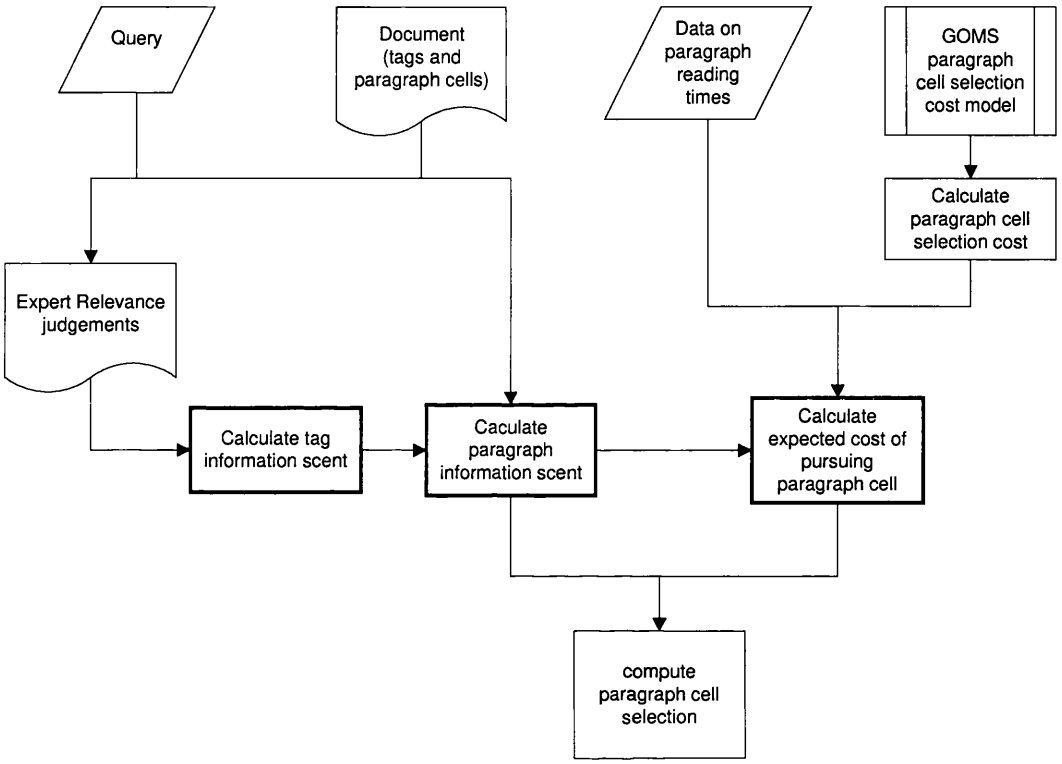


Figure 48 Overview of the paragraph cell choice portion of the performance-level best-case model

The following sections layout the details of the performance level model. Section 2.2.1 defines the measures used to quantify cost and information value. Section 2.2.2 explains how the estimation of tag and paragraph scent is modelled, and 2.2.3 explains how the estimation of cost is modelled, and how the actual performance times are calculated.

2.2.1 Currency Assumptions

Foraging models must make assumptions about what criteria are used to compare alternative values of cost and benefit; these are called currency assumptions.

Time is almost always the currency used for cost. Nevertheless, it is of course true that other factors such as anxiety, cognitive complexity and workload, are also important. Time is an important cost and is arguably a correlate of many of these other costs. Another advantage of using time as the currency for cost, is that it can be reliably predicted by theoretical models (Kieras & John 1996). Time, in seconds, is therefore be used to measure cost.

When foraging for information the benefit is obviously information itself. What is less clear is how to measure this. Pirolli and Card suggest a number of ambitious possibilities, such as memes and human capital, while themselves opting for the more pragmatic "number of relevant documents for some purpose" (Pirolli & Card 1998, p. 83). This research could have opted for a similar measure i.e. number of relevant paragraphs found. But as argued above in section 3.1.4 of chapter 5, it is important to acknowledge that 'relevant' paragraphs are not all equally relevant. To reiterate, some paragraphs will be highly relevant, containing information key to the readers goal, others will be less so.

The currency chosen for 'information' is the expert relevance rating. The judges made their judgements on a 4 point scale with 3 signifying 'key information' and 0 signifying an irrelevant paragraph (see section 3.2.3). This has the effect of placing higher value on more relevant paragraphs. Of course, it is also very crude, a paragraph of relevance 3 is unlikely to contain three times as much useful information as one with a relevance of 1. Nevertheless paragraphs with a higher relevance rating will contain more, or more valuable, information. Hence using relevance ratings as the currency, offers a pragmatic measure of information which can express the differing value of paragraphs.

2.2.2 Scent models

In order decide between tags and paragraph cells, the best-case model needs to be able to estimate the amount of valuable information to gained from each of them; in other words it needs to be able to calculate their information scent. This section is split into two, first the estimation of tag information scent is modelled in section 2.2.2.1, and then the estimation of paragraph cell information scent is modelled in section 2.2.2.2.

2.2.2.1 Tag information scent model

The tag information scent model, unlike previous information foraging models, assumes perfect information scent. The perfect information scent of an item is equal to its *actual* information value. Here the method used to calculate perfect tag scent is described and its use is justified against the background of previous models which attempt to simulate scent estimation (Pirolli & Card 1999).

Perfect tag scent is calculated for each query by simply summing up the expert relevance ratings of all the paragraphs reachable from a tag. The result is a score indicating the total amount of information obtainable from a particular tag for a particular query and document. This information scent score thus maps exactly to the amount of available information; it is thus *perfect* information scent.

Summing relevance ratings makes the invalid assumption that the relevance ratings are additive; which they are not, as they are on an ordinal scale. Nevertheless, an alternative (and more complex) scheme would involve making more assumptions that would not be any better supported.

An alternative approach would be to simulate human estimation of information scent using techniques such as those of Pirolli and Card (op. cit.). Perhaps the most serious problem with this approach is that computer simulations of semantic word relatedness

judgements, such as those using spreading activation networks, still fail to match human sophistication. For example, they would not be able to distinguish between the tags 'competitors' and 'non-competitors'. Also, due to their reliance on word frequencies the fidelity of simulated relatedness estimates are dependent on the number of words describing the items whose relatedness is being judged. For this reason their performance will be especially poor in modelling the scent of GridVis tags which are rarely more then two words long. Finally, the metadata taxonomy rests heavily on domain specific knowledge, e.g. fact that 'Tesco' is a 'competitor', which would be difficult to model using word frequency techniques[21]. Since this evaluation is attempting to look at the *best*-case performance a low fidelity simulation of scent estimation is of little interest.

On the other hand if one assumes that the tags have been well constructed then perfect information scent should give a good approximation to actual best-case human performance. The tags with the highest perfect information scent should also be those that seem the most relevant. This assumption is reasonable since the methods used to construct the metadata were chosen precisely to ensure that it was of a high quality.

Although perfect information scent should give a good approximation to the best-case human judgements, it will necessarily give a closer fit to the actual value of information available. However, since this evaluation is examining the best-case performance of GridVis it is preferable for the scent to be better, rather than worse, than that perceived by users. In this way, the evaluation can determine the upper limit on the performance a user might be able to achieve.

---

[21] This is partly due the lack of large and accessible text corpuses in this domain and due to the sublety of the domain knowledge that can be used in interpreting the taxonomy

One further question is whether the perfect scent should track the declining scent of tags as the relevant paragraphs relating to them are consumed. To see why this is important, consider two tags e.g. 'Tesco' and 'Mayfair' which both have a perfect scent of 1. The 'Tesco' tag is explored; the relevant paragraph, which is about Tesco opening a new store in Mayfair, is read. Later the user considers the 'Mayfair' tag; what should the scent of this tag be? In this case one safely can assume that an expert user will realise that this tag will probably lead to the same paragraph as the 'Tesco' tag; thus the scent would be much reduced. One approach is to ignore the user's knowledge of probable overlaps in the coverage of tags and calculate scent as the *initial information value* of a tag. This results in the model spending large proportion of its time investigating highly relevant but overlapping tags (e.g. 'Tesco' and 'big_competitor'[22]). As such the model's behaviour is far from expert. One way to model the user's knowledge of tag overlaps, is to calculate the perfect scent as the *unexploited information value* of a tag. As with perfect scent in general, this will result in scent which is more prescient than that perceived by an actual expert user. However this approach offers a tractable model which will allow an upper bound on performance to be computed.

2.2.2.2 Paragraph cell information scent model

In order to choose between paragraph cells, the best-case model needs to be able to calculate the information scent of each cell. This is complicated by the fact that four strategies for paragraph cell choice were defined in section 2.1.1. A separate scent model is needed for each of these strategies since they all differ in what they use as information scent (see Table 19). The model for each strategy will now be described in turn starting with the simplest and moving on to the most complex.

---

[22] Tesco's is the major competitor of Sainsbury's, the company at which much of this research was done

Strategy	Cues from GridVis used
Let to right (LtoR)	None
Applicability only (AppOnly)	Applicability of the cell being considered for selection.
Co-occurrence only (CoocOnly)	Relevance of other tags in column (paragraph) of interest
Co-occurrence and applicability (Cooc&RestrictedApp)	Applicability of the cell being considered for selection. Relevance of other tags in columns (paragraphs) of interest

Table 19 The different possible strategies for cell choice

The LtoR strategy, uses no information and thus assigns the same information scent to each paragraph cell; it is essentially equivalent to the commonly used 'Left to right' strategy (identified in the qualitative analysis of the initial empirical evaluation of GridVis reported in chapter 5 section 3). To model this strategy the same information value is assigned to each paragraph cell. This value is calculated by dividing the information scent of the tag being investigated by the total number of paragraph cells to which it relates. The information scent is thus evenly distributed across the paragraph's cells for the tag.

The models for the remaining strategies, require transformation of scaleless relevance values into information values. This requires a mapping function based on the relationship between relevance values and actual information values. Data for this was obtained by running the best-case model of GridVis use using perfect tag scent to determine which tags are chosen for each document query pair, in the optimal case. A listing was then obtained of all the paragraph cells on each of these tags and their information values. This data set was produced because it is only on the paragraph cells of relevant tags that the information scent will discriminate relevant paragraph cells; the applicability of an irrelevant tag to a set of paragraphs will have no bearing on the relevance of those paragraphs to the query at hand. For the sake of clarity this data set, for mapping relevance values to information values will be referred to below as 'the mapping data set'.

The AppOnly strategy only uses cell applicability in computing information scent. To model this, the relationship between a cell's applicability and its paragraph's information value was examined. The applicability of cells in the mapping data set were then listed with their corresponding information values based on the expert judge relevance ratings (see section 3.2.3 for a description of these relevance ratings). A linear regression of this data provided a function (Info_Value = applicability*0.931+0.751) that serves primarily to appropriately scale the applicability, transforming it into a information value. As can be seen from Figure 49 there is little relationship between applicability and information value. This finding is interesting in that it suggests that the applicability information included which each metadata tag is not in fact helpful in locating relevant information.

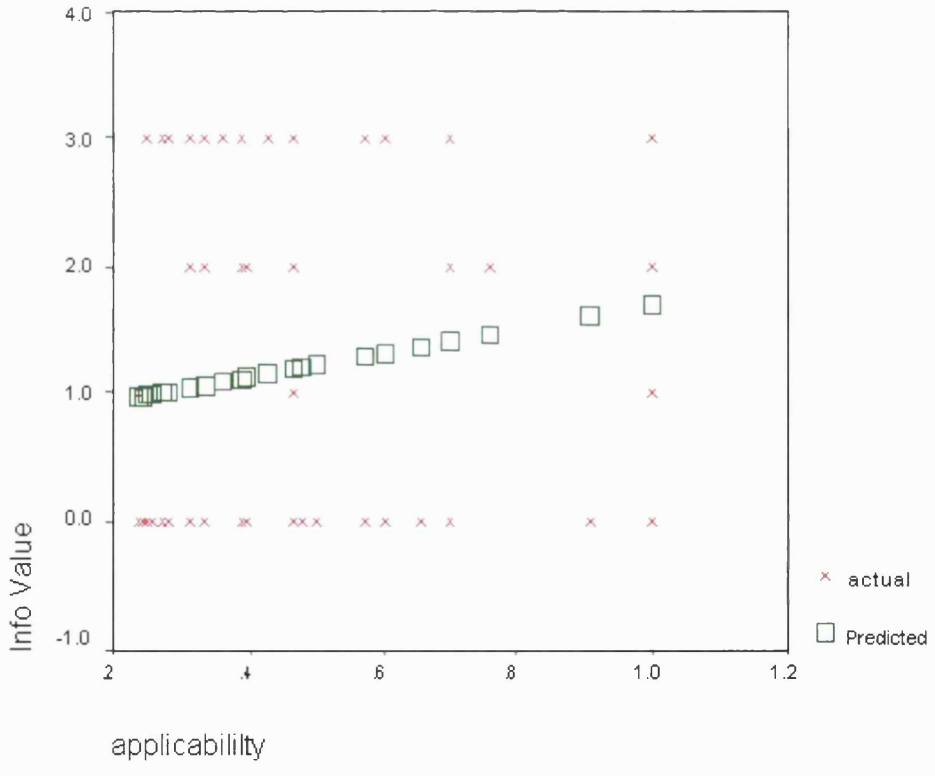


Figure 49 A scatter plot showing the correspondence between information value predictions based on applicability alone and the actual information values from the expert judge relevance ratings

The CoocOnly strategy only uses the relevance of other tags in the same column (i.e. paragraph) of interest; for the sake of brevity this will be referred to as co-occurring tag scent. To model this we examined the relationship between the sum of information scent of each co-occurring tag and the information value of the corresponding paragraph. A linear regression provided a function (Info_Value = coocurring_tag_scent*0.07813 - 0.235) with a good numerical fit to the data (R^2 = 0.618) as shown in Figure 50. This suggests that the co-occurring tags are a rich source of information scent, corroborating the conclusions of the qualitative analysis in chapter 5, section 3.3.2.

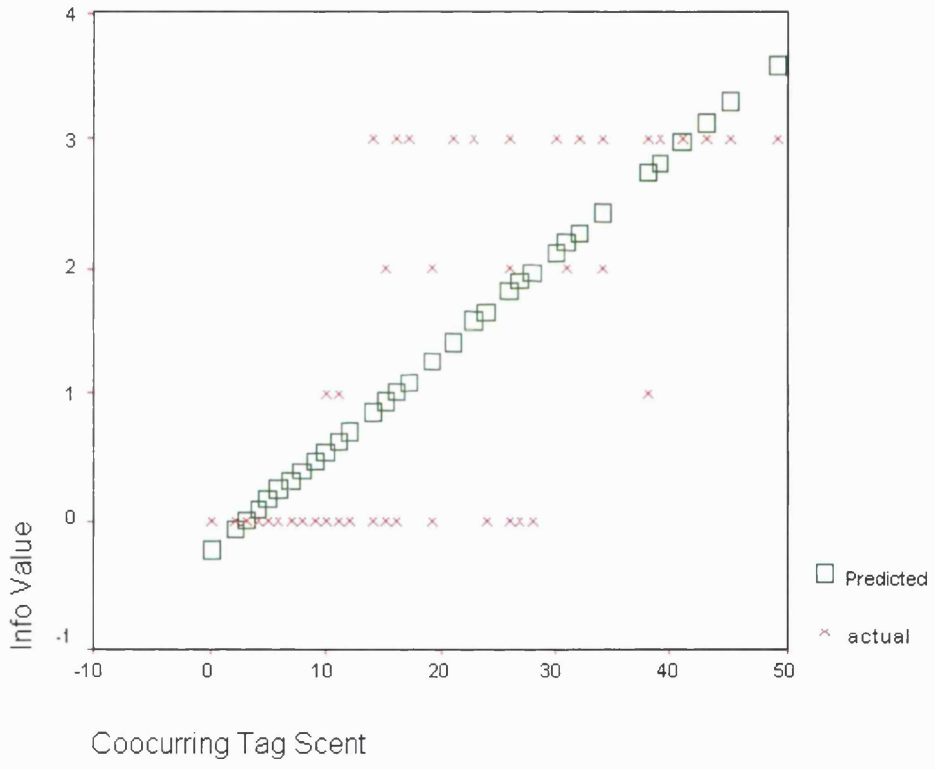


Figure 50 A scatter plot showing the correspondence between information value predictions based on co-occurrence alone and the actual information values from the expert judge relevance ratings

Cooc&RestrictedApp uses both the applicability of the cells being considered for selection and the scent of their co-occurring tags. This was modelled by assuming that the applicability is modulating the information received via the co-occurring tags. To illustrate this consider a user researching Tesco's financial position. They select the 'Tesco' tag which has two paragraph cells associated with it. Both have 'financial result' as a co-occurring tag but only first one is highly applicable to 'Tesco'. This first cell is preferable because the second might be about the financial performance of something else! So, to simply model this modulation of co-occurring tag scent by applicability we multiply the applicability (which lies between 0.5 and 1) by the sum of the co-occurring tag scents. A linear regression of the relationship between the resulting figure and info value provides a function (Info_Value = cooccurring_tag_scent* applicability*0.08086 - 0.304) (see Figure 51) with a lower, but still reasonable, numerical fit ($R^2 = 0.408$).

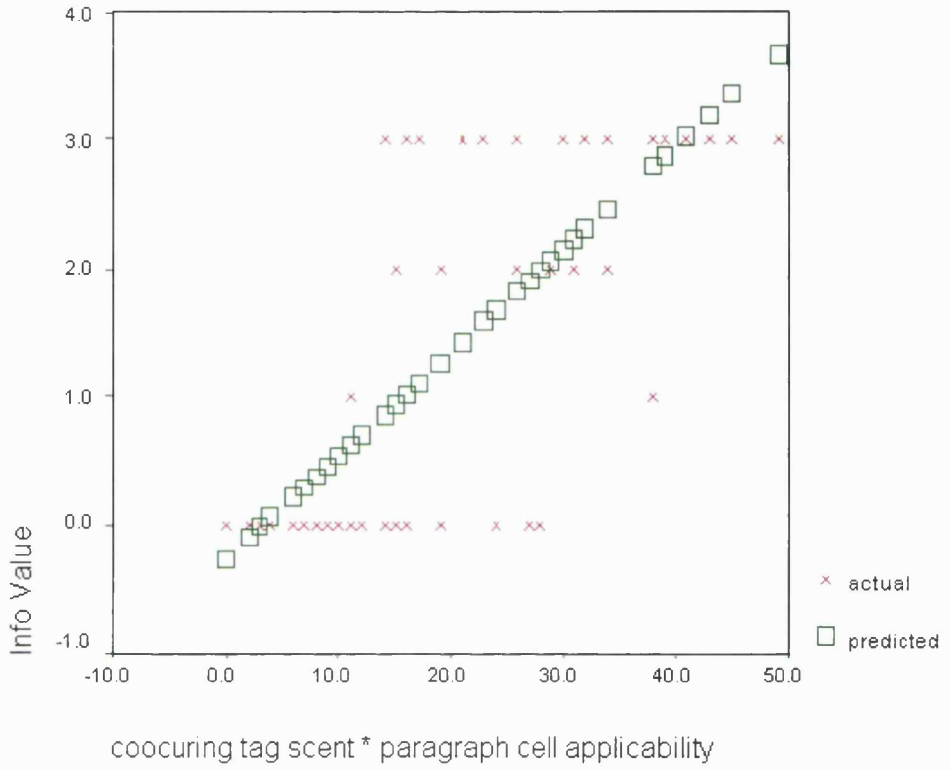


Figure 51 A scatter plot showing the correspondance between information value predictions based on applicability and co-occurrence and the actual information values from the expert judge relevance ratings

2.2.3 Cost models

Cost models are needed to compute *actual* time costs of the actions taken to calculate performance metrics and *expected time costs* of alternative potential interaction paths are needed to make foraging decisions.

2.2.3.1 Models of actual cost

The *actual* costs involved in using GridVis were predicted by three separate models. These models predict the tag choice cost, the paragraph cell choice cost and the paragraph handling time respectively. The way these are broken down, and the data used to parameterise each component, are shown in Figure 52.

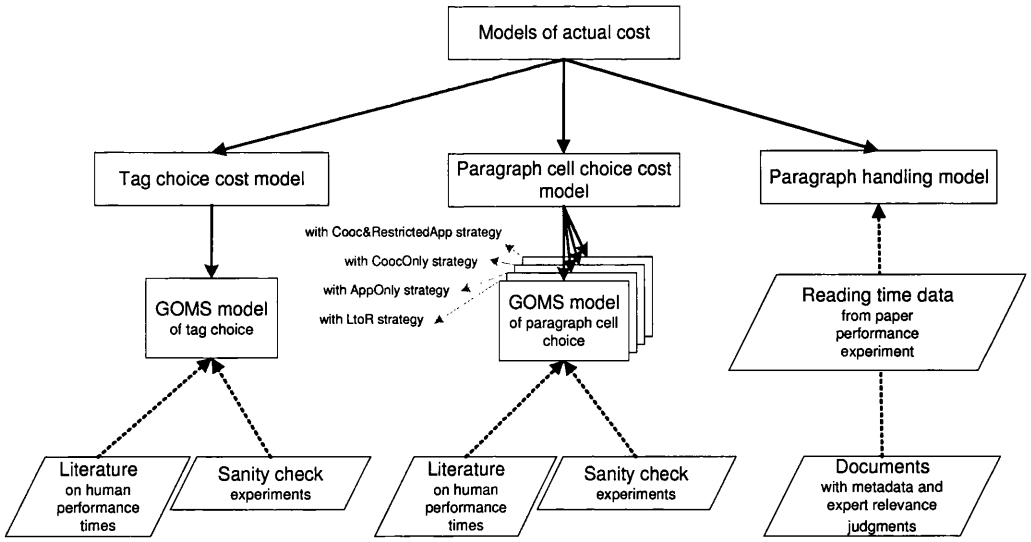


Figure 52 An illustration of how the actual cost models, ▭, are broken down, and what data, ▱, is used to set their parameters

The tag choice and paragraph cell choice cost models are both primarily NGOMSL models with an information foraging theory control structure in place of selection rules. The NGOMSL technique was used because it provided a well defined and documented technique for developing complex models. A lower level GOMS model such as CPM-GOMS (Gray, John & Atwood 1993) was not used since its higher accuracy was expected to be insignificant next to the error introduced by the approximate nature of the strategy level model.

The GOMS models were developed from the expert strategies. Although the expert strategies had already been defined, they needed to be worked out at a lower level of detail for the GOMS models. This was done by iteratively rewriting the models, following Kieras' guidelines (Kieras 1994). On each iteration, they were simplified, debugged and streamlined. The resulting models were then coded in Visual Basic, which proved very useful in identifying inconsistencies and further improvements. The performance predictions from these models were then compared to human performance data from a simple 'sanity check' experiment.

Paragraph handling time was modelled by deriving simple linear functions, using linear regression, from the data obtained in the paper performance experiment (see section 3).

2.2.3.1.1 *Tag choice cost model*

The tag choice cost model produces numerical predictions of the time taken to choose a given tag for a given document. It is a GOMS model with an information foraging theory control structure. The GOMS model will only be described at a high level here with a few pertinent issues drawn out for further analysis. The full model can be found in Appendix 4.

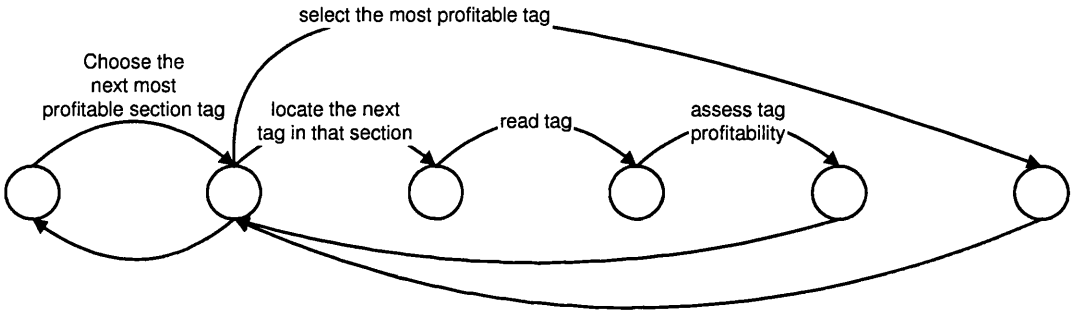


Figure 53 The basic structure of the tag choice NGOMSL model

The basic structure of the NGOMSL model is shown in Figure 53. Each of the circles represents a state, and each arrow represents a stage in the model. First, the model has to choose a section, then it has to choose a tag from within that section. To choose the most profitable tag in the section the user has to read and assess each one. Once all the tags have been read the most profitable is chosen and selected. The total time is calculated as the sum all the operator times invoked by these steps in the NGOMSL model.

When the model is done exploiting the tag, it returns to the taxonomy section just inspected to consider whether there are any more tags worth selecting. If there are no more worthwhile tags in that section (as computed by the optimal diet selection algorithm layed out in section 2.1), the next most profitable section is examined.

In the 'assess tag profitability' step of Figure 53 the model needs to estimate both a tag's information scent and the expected cost of exploiting it. To assess the information scent, the tag is examined and compared with the query. The cost is assessed by looking along the tag's row to estimate how many paragraph cells there are and how wide they are (column width is proportional to paragraph length). This information about the paragraph's cells can then be used with the tag's information scent to estimate how much text will need to be read. This is dealt with in more detail below in section 2.2.3.2. Thus for each tag an estimate of profitability is built on estimates of information scent and cost.

The model assumes that the cost of choosing a tag from a section in the taxonomy decreases monotonically each time the section is revisited. One would expect that when a user returns to a section in the taxonomy they will have some memory of the section's information scent. This memory will make choosing the next tag faster. Also, the more times the section has been revisited the better the user's memory will be and therefore the faster their choices. To model this, the costs were discounted using the simple function given below as Equation 4 and illustrated in Figure 54. This function clearly over-simplifies the effect of memory on item choice performance, but in the absence of a principled understanding, a simple model is preferable to a more complex and equally unfounded one.

$discount = \frac{(n-k)}{n}$, where n is the number of tags in the section, and k is the number of times the section has been visited

Equation 4 A simple function used to discount the costs from memory dependent portions of the GOMS model

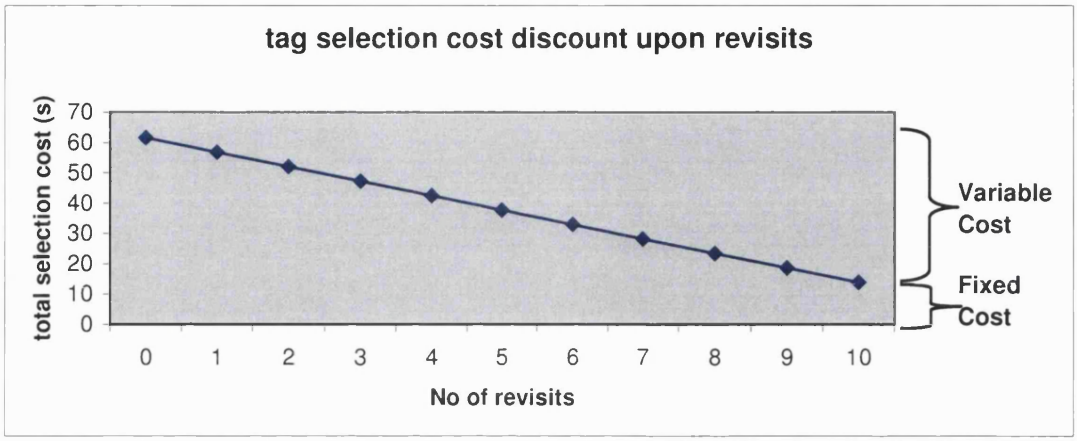


Figure 54 Chart illustrating the discounting of tag choice cost each time a section is revisited

The detail of the operators chosen and of the GOMS model's design is perhaps best absorbed by inspecting the model itself, which is given in Appendix 4. Some of issues arising from the choice of times for operators particular to this model, will now be discussed, and the complete set of operators are shown in Table 20 and Table 21.

M, the mental operator, is used for a number of different high level operators. This is a standard procedure in the construction of NGOMSL models (Kieras 1994). Broadly, there were three types of operator dealt with in this way: 'complex evaluations based on working memory contents', 'memory retrievals with simple evaluations' and 'complex perceptual cognitive operations'. The first two classes of operator are essentially memory based. This makes M appropriate for them since the time value associated with it, 1200ms, is based on experimental measurements of tasks requiring memory based

mental activity (Olson & Olson 1990). Nevertheless, investigations were made into the psychological literature to find a more accurate time for the 'compare to query' operator. Times were found for the closely related but not identical tasks of lexical decision and thematic inference (see Table 20); these yielded times which were very similar to M. Because of this, and for the sake for consistency, M was used instead of these times. M was also used for the single 'complex perceptual cognitive operation'. This was done, basically, for lack of better alternatives; the operator is unique to GridVis and is simply too complex to break down into primitive operators.

High level operators	Execution times
Locate <tag >	230ms (MHP eye movement (Card, Moran & Newell 1983))
Get-from-task <next tag>	340ms (John & Newell 1989)
Are no more tags in <section>	430ms 230ms+100ms+100ms (model human processor eye movement, and perceptual processor (Card, Moran & Newell 1983), cognitive cycle (Bovair, Kieras & Polson 1990))
Compare query	1200s M –Analyst defined working memory based mental operator (Olson & Olson 1990) (Lexical decision ~1.2s (Collins 1969) and thematic inference judgements 1.3s for yes and 2s for no (Reder 1979)
Choose unexamined <tag> with highest info scent	1200s M – Complex evaluation based on working memory contents -> analyst defined mental operator (Olson & Olson 1990)
Calculate profitability	
<tag> is not worth inspecting	
Have just been investigating a tag from <section>	1200s M – Memory retrievals with simple evaluations -> working memory based mental operator (Olson & Olson 1990)
Retrieve <last examined tag> from LTM	
Have just selected a tag in <section>	
Tag already examined	
Remember and home in on most relevant un-examined subsection	
Get-from-task <number of cells on next tag><total width of cells on next tag>	1200s M – Complex perceptual cognitive operation - Analyst defined mental operator (Olson & Olson 1990)

Table 20 High level operators for the tag choice GOMS model

Primitive operators	Execution times
store info scent in WM Store in number and width WM	0ms – for expert working memory operations (Kieras 1994)
locate and point to tag	1100ms (Card, Moran & Newell 1983)
click tag	200ms (Card, Moran & Newell 1983)
verify click	0ms – for experts verification operations (Kieras 1994)

Table 21 Primitive operators for the tag choice GOMS model

2.2.3.1.1.1 Sanity check experiment

For the purposes of the best-case evaluation the GOMS model need only provide a reasonable fit to the way actual performance times vary according the features of the environments such as the number of tags in a section, or the number of times a tag has been revisited. Nevertheless it would be unacceptable if the GOMS model gave predictions that were systematically very different from human tag choice times. So, in order to sanity check the GOMS model a small experiment was carried out to provide data on human tag choice times.

This experiment used an instrumented version of GridVis (described in section 2.1.2) to time the tag choices. The experiment only used one participant (the author of this thesis) who had internalised the 'best-case tag choice strategy' and was very familiar with the tag hierarchy. This experiment used five of the queries developed for the best-case evaluation (see section 3.2.2). For each query a number of relevant sections were noted down. Twenty five trials consisted of choosing a tag from a pre-designated section. The trial was begun by looking up which query and section were to be used. Then the start button was selected, revealing the metadata tree, and the participant read the tags in the specified section and selected the tag judged to be the most relevant. Then the participant read some unrelated material, as a short distracter task, before moving on to the next trial. The distracter task, which consisted of doing other research work for

about two minutes, was included to mimic the effect of exploiting the tags paragraph cells.

The results of this experiment are displayed in the graph below (Figure 55). The absolute mean percentage difference between the GOMS predictions and the human tag choice time was 5%, which is within the recommended maximum 20% (John & Newell 1990). This suggests that the GOMS model, and the algorithm and operators it is made up of, is indeed doing roughly the right thing.

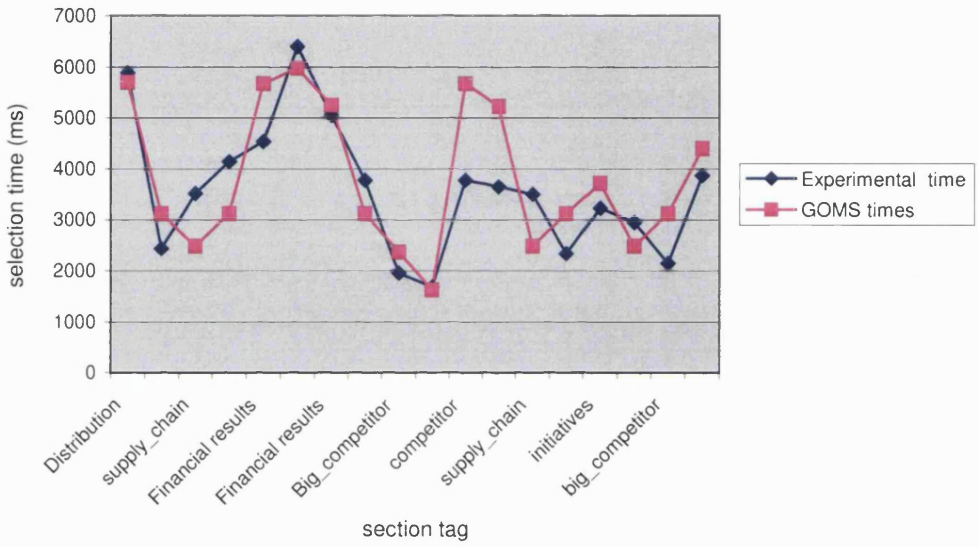


Figure 55 Human and GOMS times for tags choice. The trials are shown in sequential order; thus, repeated section tags indicate section revisits

2.2.3.1.2 *Paragraph cell choice cost model*

The paragraph cell choice cost model provides a prediction of the time taken to choose a given paragraph cell. The time taken to do this is needed to calculate the task time performance metric. As with the tag choice cost model, (section 2.2.3.1.1), the model used is essentially a NGOMSL model with a information foraging theory control structure. The full details of the model are available in Appendix 4. Here a high level

description of the model will be given, with important points being drawn out for discussion.

Since, there are important differences in the alternative strategies for paragraph cell choice, separate NGOMSL models were constructed for each one. As described above (section 2.1.1) the strategies differ in the amount of scent they use (Table 19). As such they can be ordered in terms of increasing complexity, with the simpler models acting as foundations for the more complex ones.

We will start with the LtoR strategy, which is simplest since it makes no attempt to gather information scent. The GOMS model for this strategy, illustrated in Figure 56, only actually consists of one method, the user scans the row corresponding to the tag of interest, from the left hand side, and selects the next, unvisited but shaded (i.e. applicable), paragraph cell.

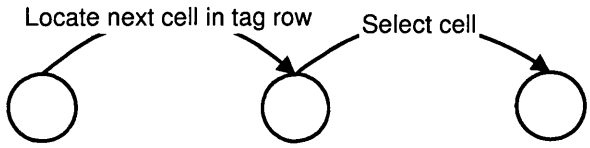


Figure 56 Illustration of the cost model for the LtoR strategy

The AppOnly strategy involves making a minimal amount of effort to assess the information scent of each cell on the row of interest. In the GOMS model for this strategy, illustrated in Figure 57, a single visual search isolates the row's brightest cells, this cell's shading indicates the applicability and thus dictates the information scent and profitability. The brightest and therefore the most profitable cell is selected.

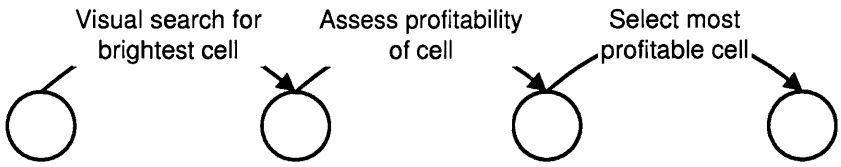


Figure 57 Ilustration of the cost model for the AppOnly strategy

Once the selected paragraph has been dealt with, the model returns to that row to see whether there are any other cells worthwhile selecting. One would expect the memory from the previous cell choice to speed process of choosing the next one. For this reason the costs of re-inspecting the row is discounted with repeated visits. This is done using Equation 4 as illustrated in Figure 54 and explained earlier.

The cost model for the CoocOnly strategy[23] is illustrated in Figure 58. In this strategy the scent of a paragraph is inferred from that of the tags that have been applied to it, i.e. the co-occurring tag scent. The model searches for paragraph cells on the row of interest. For each paragraph cell the model reads the label of each co-occurring tag, and uses this to estimate the cell's information scent and thence its profitability.

---

[23] The construction of a GOMS model for the CoocOnly strategy lead to an interesting insight into the detail of the best-case strategy which has implications for GridVis's design. In the first iteration of this model every highlighted tag was read. This takes advantage of the fact that every tag applied to the paragraph currently under the mouse is highlighted. The problem with this is that every section with highlighted leaf tags in it is also highlighted. This means that generally about 50% of highlighted tags are section tags, and these are much less informative than leaf tags. In addition, while picking out highlighted tags only requires a simple pre-attentive visual search (Healey, Booth & Enns 1996), reading tag labels and assessing their information scent is relatively very costly. A much more time efficient strategy is to use a slower serial search to pick out half the number of tags i.e. only the non-section tags. A revised version of the coocOnly model used this more efficient strategy.

This insight from the GOMS model suggests that it is important to clearly distinguish co-occurring section tags from non-section tags. As a result, GridVis was changed so that co-occurring section tags were assigned a light grey highlight as opposed to the green and orange ones given to non-section tags.

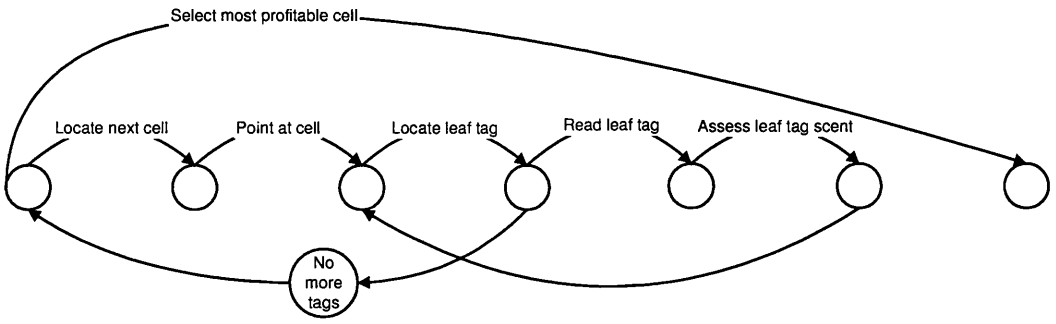


Figure 58 Ilustration of the cost model for the CoocOnly strategy

The Cooc&RestrictedApp strategy, illustrated in Figure 59, is essentially a combination of the CoocOnly and AppOnly strategies. It makes use of both the cell's applicability and its co-occurring tags to estimate profitability.

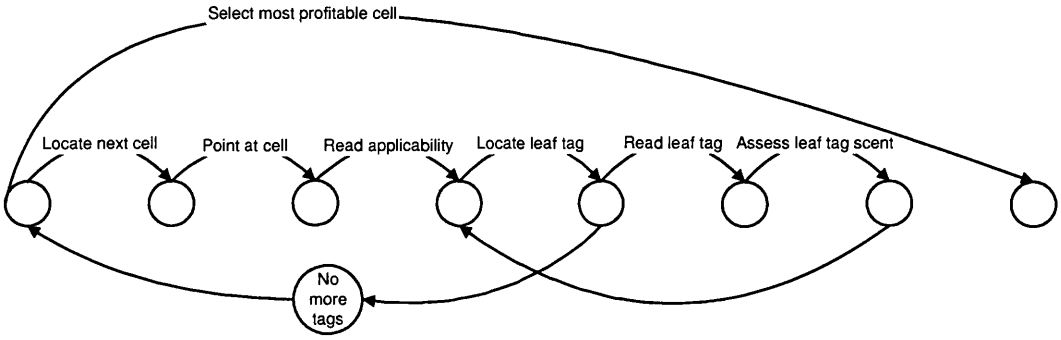


Figure 59 Ilustration of the cost model for the Cooc&RestrictedApp strategy

The detail of the NGOMSL models and the operators that are used within them can best be understood by studying the full GOMS models in the Appendix 4. However, the full list of operators are presented below in Table 22.

Two points are worth bringing out at this stage. Investigations into the perceptual psychology literature uncovered data for parallel and serial visual search times (Healy 1996). This was used to provide operators (or rather functions) sensitive to the difficulty of the search task, (i.e. the number of cells in the row). Also, the 'get-from-task <cell-applicability>' operator was built up, using a established procedure (John & Newell

1990), from the perceptual and cognitive cycle operators of the Model Human Processor (Card, Moran & Newell 1983) using principles from information theory (see Appendix 4).

High level operators	Execution times
Choose cell with highest information scent	1200ms M - Complex working memory based cognitive operator - Analyst defined mental operator – (Olson & Olson 1990)
Visual search for unselected cell on <tag> row with highest information scent	0 ms (since always followed by a point– (Gong 1993)& (Kieras 1994)) Otherwise, it would be M - 1200ms (i.e. same as *locate* operation (John & Kieras 1996))
Primitive operators	Execution times
[Is] there is another highlighted leaf-tag	150 ms (perceptual operator (100ms) + cognitive operator (50ms), from MHP (Card, Moran & Newell 1983)) – this operation is acting on a perceptual memory of where the highlighted tags are.
Locate <highlighted tag >	230ms (MHP eye movement (Card, Moran & Newell 1983)) – this is used when the location of the tag is already known
Visual search <highlighted leaf tags>	652ms Parallel colour search (500ms to 750ms (Healey, Booth & Enns 1996)) – since colour searches (i.e. filled vs. not filled) can be pre-attentively conjoined with spatial location (Ware 2000) OR 0ms when followed by a point operation (Gong 1993)
Visual search for next cell on <tag> row	0 ms (since always followed by a point without an intervening decision – (Gong 1993; Kieras 1994)) Otherwise it would be 652ms Parallel colour search (500ms to 750ms (Healey, Booth & Enns 1996)) – since colour searches can be pre-attentively conjoined with spatial location (Ware 2000)
Visual search for <last selected cell>	~760 + #cells_on_row*8 ms (Healey, Booth & Enns 1996) – taken from graphs of serial search of conjoined visual features– assumes that spatial conjoining is still pre-attentive for serial colour search OR 0 ms (when the search is followed by a point (Gong 1993))
Visual search for next unselected cell on <tag> row finds a cell	
Visual search for first unselected cell on <tag> row finds a cell Visual search for first unselected cell on <tag> row	

Visual search for first <highlighted leaf tag> Visual search for next <highlighted leaf tag>	
Get-from-task < highlighted tag>	340ms (John & Newell 1989)
Get-from-task <cell-applicability>	Tp+Tm*2.7 235 ms see Appendix 4
Store information scent of <cell> in WM	0 ms – for experts WM operations (Kieras 1994)
Update information scent estimate in WM	0 ms – for experts WM operations (Kieras 1994)
Verify selection Verify mouse position	0 ms– for experts verification operations (Kieras 1994)
Point at <cell>	1100 ms (Card, Moran & Newell 1983)or 1160 ms using Fitts Law (assuming distance 1cm, target width 0.5)
Select <cell>	200 ms (Card, Moran & Newell 1983)

Table 22 Operators for the paragraph cell choice GOMS models

2.2.3.1.2.1 Sanity check experiments

The models above are only intended to approximate actual performance, yet it is important to verify that they are indeed making sufficiently accurate predictions. To this end, four sanity check experiments were carried out, one for each of the cell choice strategies. These experiments only had one participant (the author) who had internalised the paragraph cell choice strategies and had a good understanding queries and tags.

The data from these experiments acted as a guide, roughly indicating what the GOMS predictions should be. As such they revealed that all but the LtoR model overestimated the times needed. This prompted a re-examination of the AppOnly GOMS model and the discovery of a more efficient strategy. This model gave a much better fit to the data. Similarly a single improvement to the strategy for the CoocOnly and Cooc&RestrictedApp models also gave a more satisfatory fit to the data. This improvement still left a substantial difference between the human and GOMS times. However, further changes could not be made without compromising the logic of the model. It is unclear whether the difference between the human and GOMS times is the result of sophisticated human strategies which cannot be modeled with NGOMSL, or

due to the overfamilarity of the participant with the materials. The strategies, described above, and the results report below are those for the final versions of the GOMS models.

The same basic methodology was followed for all four experiments; they all used an instrumented version of GridVis, with a pre-prepared list of paragraph selection tasks, to repeatedly enact each of the different paragraph cell choice strategies. An instrumented version of GridVis was used so as to record the cell choice times. In this version of GridVis the grid was hidden until a start button was pressed. When the start button was pressed the timer is started and then once a cell is selected the grid is once again hidden and the choice time is recorded. The cell choice tasks were specified with a document, query and tag. A list of these was compiled from the documents, queries and tags used in the sanity check experiment for the tag choice GOMS model (reported in section 2.2.3.1.1.1). For each experiment, every trial starts by reading the query, and selecting the tag specified for that trial. The start button is then pressed and an as yet unselected cell, is chosen and selected. The text produced is then read (as a distractor task) and then the next worthwhile cell on the tag is chosen and selected. When there are no more worthwhile cells left on the tag, the next document, query and tag combination in the list is used. Trials where errors were made were discarded.

The first experiment examined the time taken to choose cells using the LtoR strategy. The absolute percentage difference between the GOMS predictions and the human cells selection time was 3%, which is within the recommended limit of 20% error (John & Newell 1990). So although this GOMS model does not model the variation in human performance (see Figure 60) overall it is indeed doing roughly the right thing.

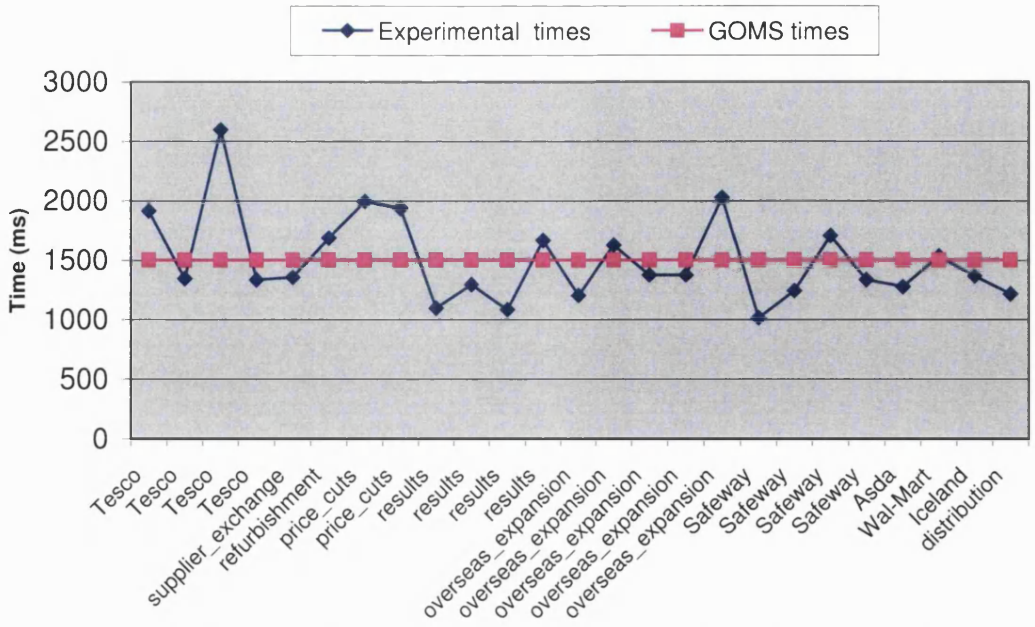


Figure 60 Human and GOMS cell choice times using the LtoR strategy

The next experiment looked at the AppOnly strategy. Here the absolute percentage difference between the GOMS predictions and the data is 15% again within the recommended limit of 20% error (John & Newell 1990). Again although the model fails to capture the variation (see Figure 60) in performance times it does predict performance times that are roughly correct.

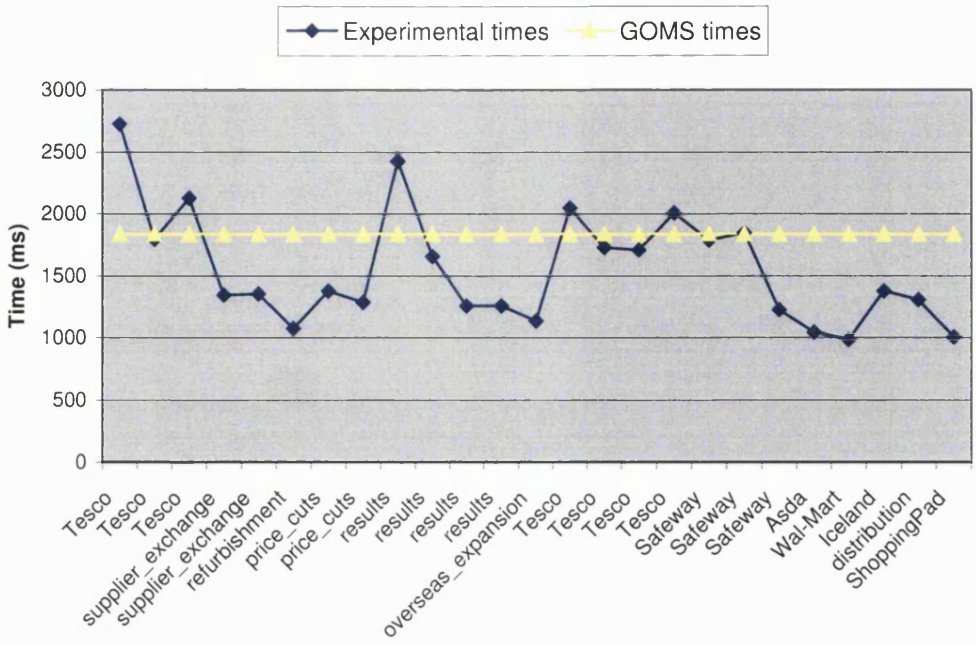


Figure 61 Human and GOMS cell choice times using the AppOnly strategy

The experiment for the CoocOnly strategy produced rather different results; the model captured the variation in the tag choice times (see Figure 62) but consistently overestimated choice times. So although the correlation between the two is high (r=.86), the mean absolute percentage difference between the experimental times and the GOMS predictions is 46%. This is probably due to the many subtle strategies that were adopted by the experimenter *even* though he was consciously trying to follow prescribed strategy. Research has shown that people make non-deliberate changes to low level 'microstrategies' in response to task structure that result in savings of a few milliseconds (Gray & Boehm-Davis 2000). For example the time taken to read tags may be subtly varied: tags in less relevant sections may only be scanned briefly; tags on paragraphs about the same irrelevant topic as previous paragraph maybe scanned more

briefly. Since the participant had constructed both the queries and the documents, his familiarity with the material could have resulted in more of these kinds of saving than would be expected from other 'experts'.

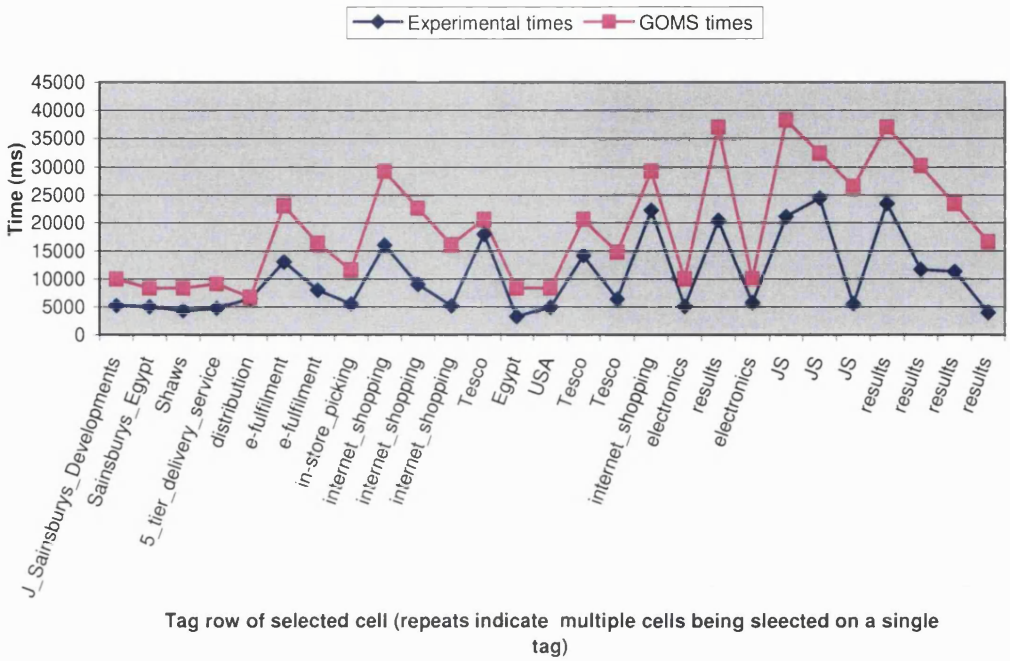


Figure 62 Human and GOMS cell choice times using the CoocOnly strategy

The final experiment looked at the Cooc&RestrictedApp strategy. As with the CoocOnly strategy the model seems to capture the variation in choice times ($r = 0.74$), but consistently overestimated them (absolute percentage difference 57%). Due to the similarity of the Cooc&RestrictedApp and CoocOnly strategies the same subtle sophistications in low-level human strategy, possibly fuelled by high familiarity with the materials, could explain the GOMS models inaccuracy.

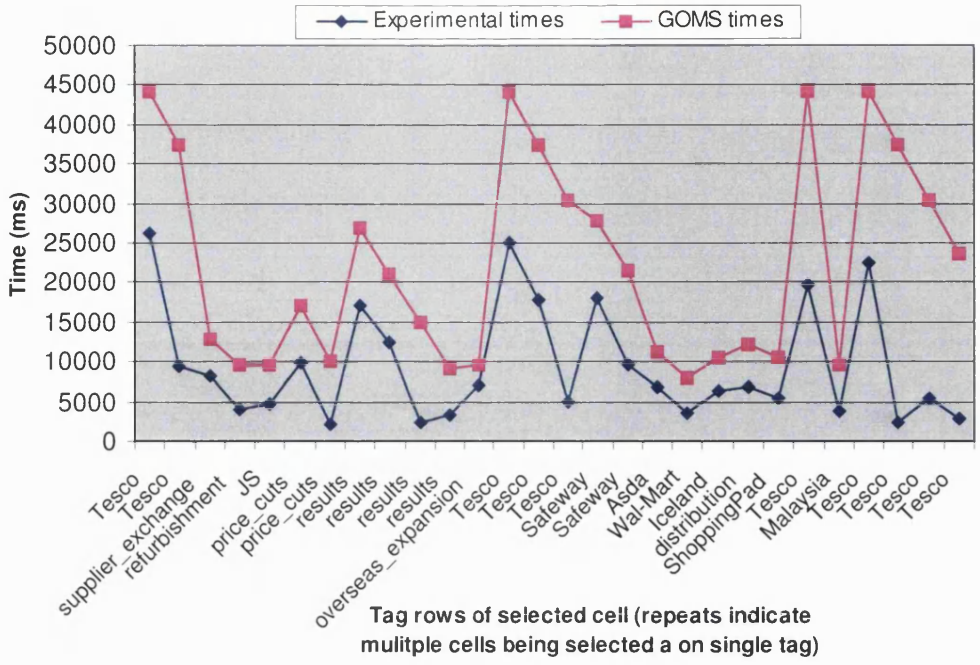


Figure 63 Human and GOMS cell choice times using the Cooc&RestrictedApp strategy

2.2.3.1.3 *Paragraph handling model*

In order to compute the total task time the model needs to calculate the actual time that is spent reading; these calculations use all the information available to produce a calculation which is as accurate as possible.

Once users have selected an instance, they must then read the resulting paragraph, giving it enough time to extract any relevant information or rapidly abandoning it when it proves irrelevant. For the purposes of the best-case model it is not necessary to model the subtle details of this process. All that is needed is a model which produces estimates of cost that are roughly equivalent to the paragraph handling costs observed in the paper performance experiment described in section 3 below. This avoids any unwanted variables introducing systematic differences between reading costs in the benchmark and GridVis.

The paragraph handling times were all calculated by using linear regression of the data from the paper performance experiments to produce simple linear functions. The data used was composed of the individual paragraph handling times of sixteen participants together with corresponding paragraph information values and paragraph lengths. In order to eliminate unrepresentative data points, the paragraph times which could not be clearly determined during transcription from video data (see Appendix 5) were omitted.

The paragraph handling time model, needed to do two separate regressions, one for accepted paragraphs and one for rejected paragraphs. This is because rejected paragraphs are read in a very different way from accepted ones. While accepted ones are read in some detail and then highlighted, rejected ones are abandoned after being rapidly scanned. The times taken with rejected paragraphs are therefore much lower than those with accepted ones so that the times from accepted and rejected paragraphs cannot be modelled with a single linear function. Hence, two separate regression equations were produced.

For accepted paragraphs a stepwise regression was performed with paragraph time as the dependent variable and paragraph length (i.e. number of words - noWords) and information value (i.e. its relevance rating - relevance), as independent variables. This produced a function (paraTime = noWords*116.919 + relevance*951.996 + 4483.907) with a reasonable (R^2 = 0.275) and significant ($F(2,449)=85$ $p<0.001$) numerical fit. Most importantly, this function serves to provide a simple summary of the inevitably noisy data (see Figure 64 and Figure 65).

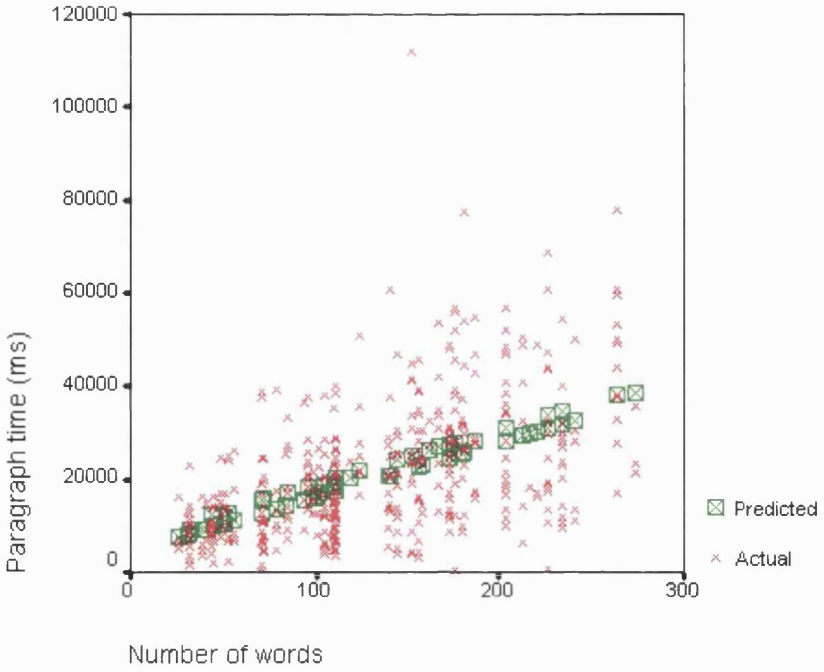


Figure 64 Scatter plot of paragraph length against predicted and actual times (from the paper performance experiment) for paragraphs that are ultimately accepted

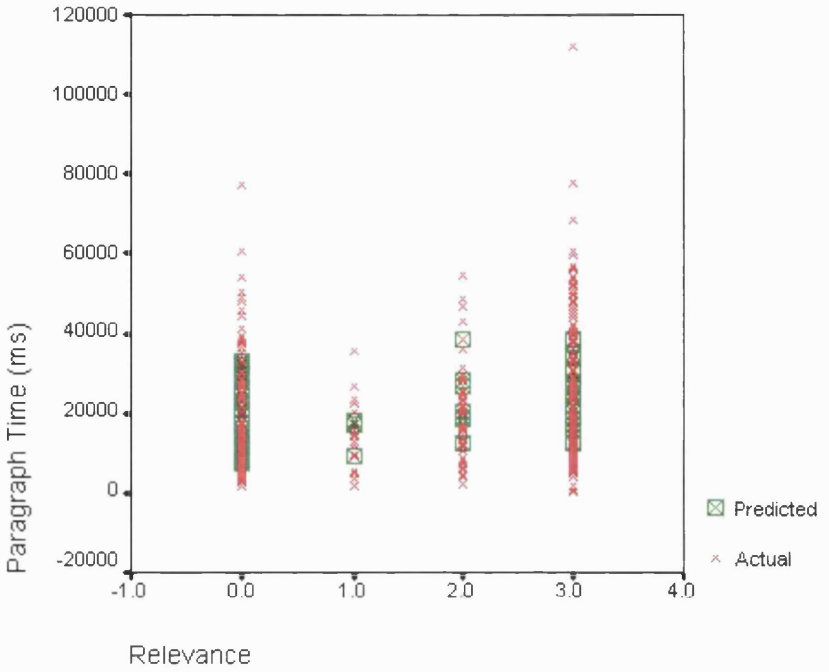


Figure 65 Scatter plot of paragraph relevance rating (i.e. information value) against predicted and actual times (from the paper performance experiment) for paragraphs that are ultimately accepted

The same regression was performed for the rejected paragraphs. The regression produced a function (paraTime = noWords*15.045 + relevance*2128.401+ 1239.530) a rather low (R^2 = 0.136) yet significant ($F(2,3225)=252.765$ $p<0.001$) numerical fit. Interesting here, the paragraph length (i.e. word no) has a much smaller effect; this is consistent with the initial use of skimming strategies which are relatively independent of paragraph length (see results from the paper performance experiment in chapter 7 section 1.1). The graph for this regression, shown in Figure 66, illustrates how the regression equation again serves to provide a simple summary of the inevitably noisy data.

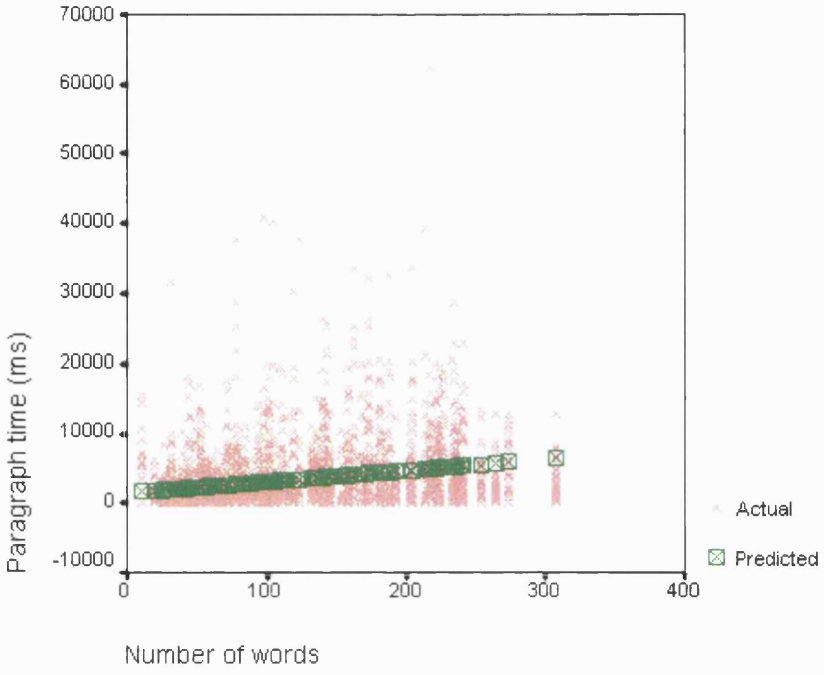


Figure 66 Scatter plot of paragraph length against actual and predicted times for reading paragraphs that are utlimately rejected

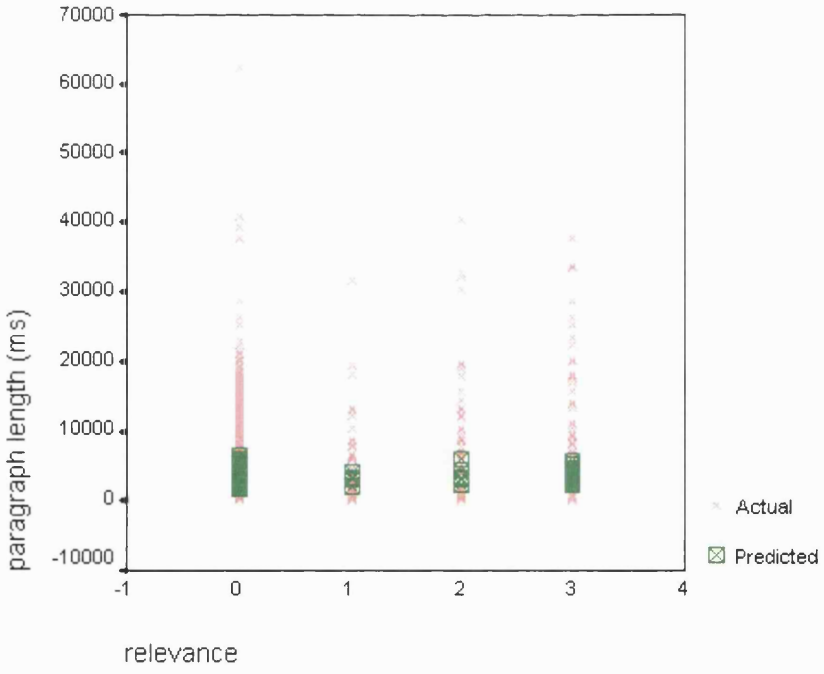


Figure 67 Scatter plot of paragraph relevance (i.e. information value) against actual and predicted times for reading paragraphs that are ultimately rejected

The 'actual handling time' model is thus given by equation 5:

If paragraph information value = 0

paraTime = numWords*15.045 + relevance*2128.401+ .239.53

else

paraTime = numWords*116.919 + relevance*951.996 + 4483.907

Equation 5 Handling time (in seconds)

2.2.3.2 Models of expected cost

The expected cost of pursuing an interaction path partly determines whether a forager will chose it. Thus the best-case model must simulate the estimation of exploitation costs. The models described in this section allow the calculation of expected costs based on the data generated by use of the models of actual cost described above in section 2.2.3.1 and the three documents prepared for use in the this best-case analysis, as described in section 3.2.1. This data is used to mimic an expert user's knowledge about the time it takes to perform operations with GridVis and read the resulting document text.

Information foraging decisions require the estimation of two broadly different types of expected costs, namely exploitation costs and average patch choice time. The estimated cost of exploiting a given section, tag, or paragraph cell is needed in the cost/benefit trade-offs that determine information foraging decisions. The essential difference between these *expected* cost models and the *actual* cost models is that the estimates of expected cost must be based solely on the information given in the GridVis interface and on the past experience a best-case user can be expected to have. As illustrated in Figure 68, there are essentially two distinct portions in these expected cost models, there

is a estimate of the amount of time spent using GridVis and there is one of how much time is to be spent handling paragraphs, i.e. reading and processing them.

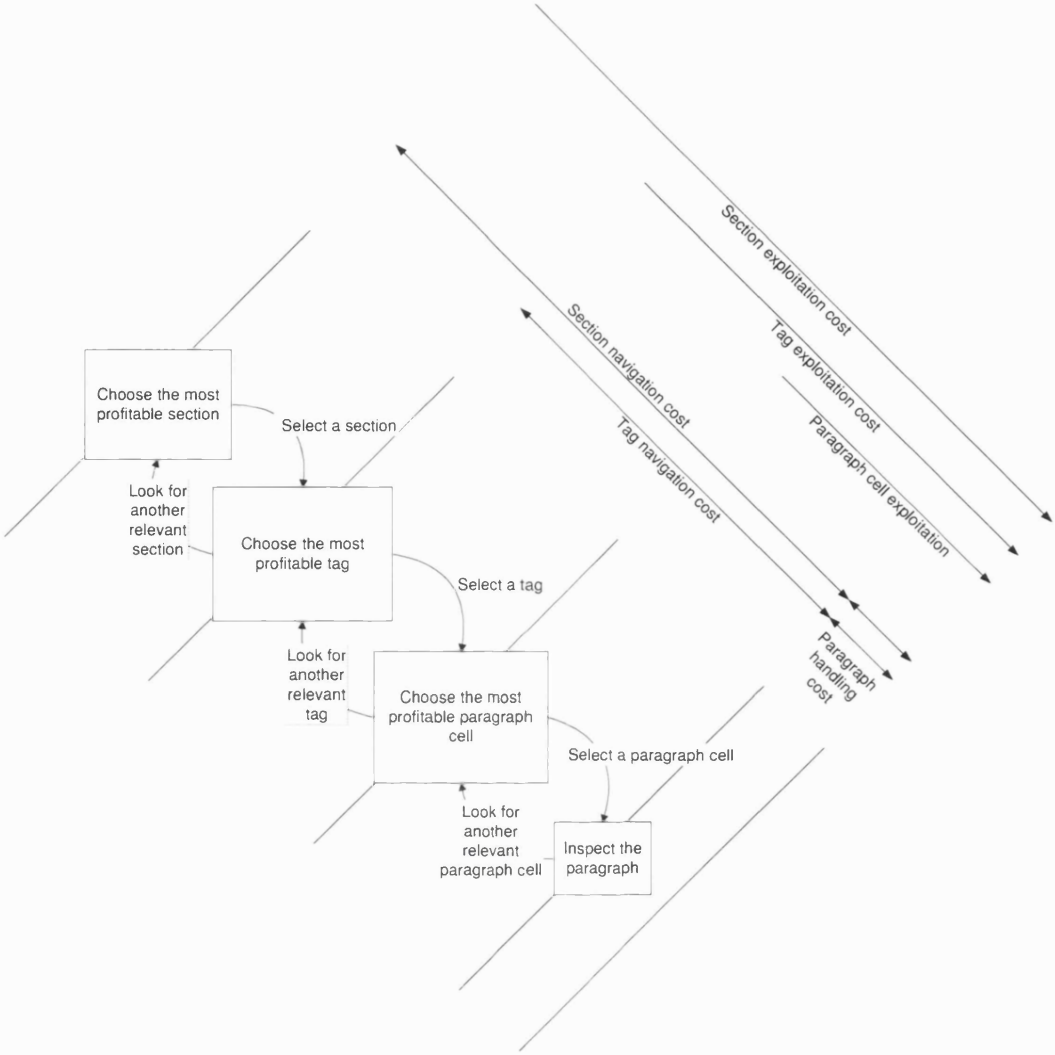


Figure 68 Illustration of the different processes involved in section and tag exploitation cost

How the estimates of GridVis navigation time and paragraph handling are modelled make up the next two sections (sections 2.2.3.2.1 and 2.2.3.2.2). Finally section 2.2.3.2.3 describes the estimation of the average patch choice time since it is used as part of the diet selection algorithm. The way the models in each of these sections relate to each other, and the data that is used to parameterise them, in shown in Figure 69.

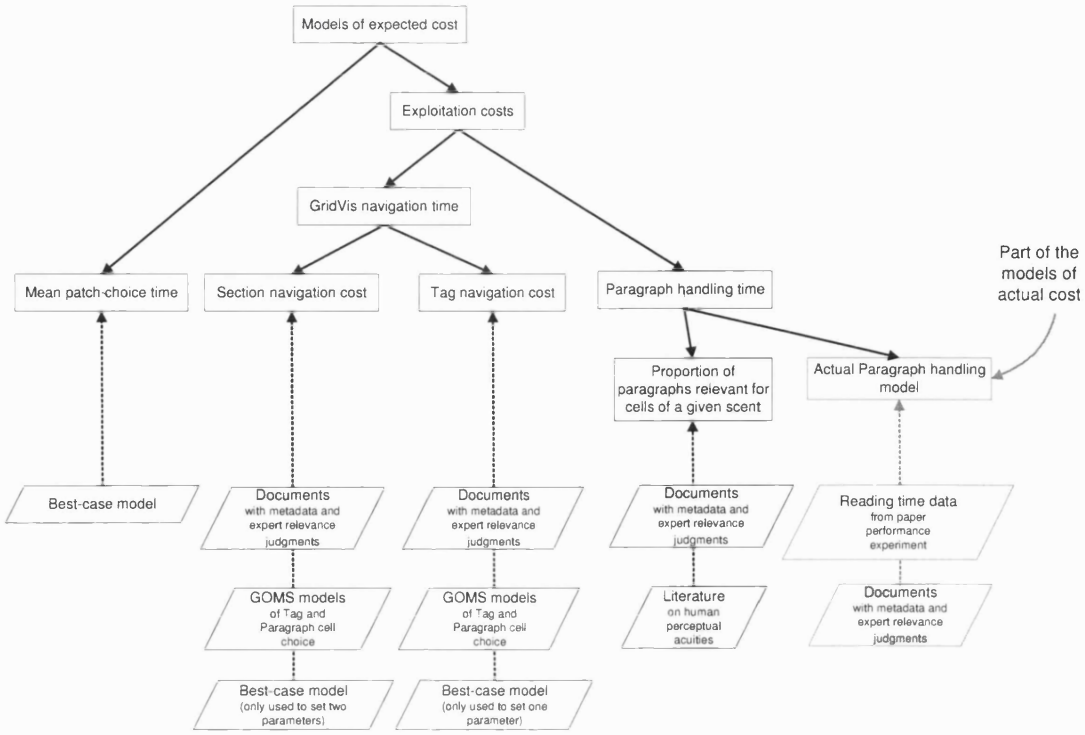


Figure 69 An illustration of the components of expected cost, ▭, and what data, ▱, is used to set their parameters

2.2.3.2.1 *Estimating GridVis navigation time*

The expected navigation time must be estimated at two points: when tags are chosen and when sections are chosen. An expert user is assumed to be able to roughly estimate the navigation time based on information available from the GridVis interface. To model these estimates the process, shown in Figure 68, was broken down into portions which can be reliably estimated. This was first done for tag exploitation, and then for section exploitation. Since section exploitation is partially a function of tag exploitation, the modelling of tag exploitation will be described first.

For tags this navigation cost can be broken down as follows: the tag must be selected, one or more paragraph cells must be chosen, and then these must be selected; this break down is shown in Equation 6. Each of these components can then be modelled or estimated, in a way that attempts to mimic the experience of an expert user, by relying

on data from the documents and the GOMS models. The way each of these components is broken down is illustrated in Figure 70.

TagNavigation_Cost = TagSelection_Cost + AllCellChoices_Cost +
AllCellSelections_Cost

Equation 6

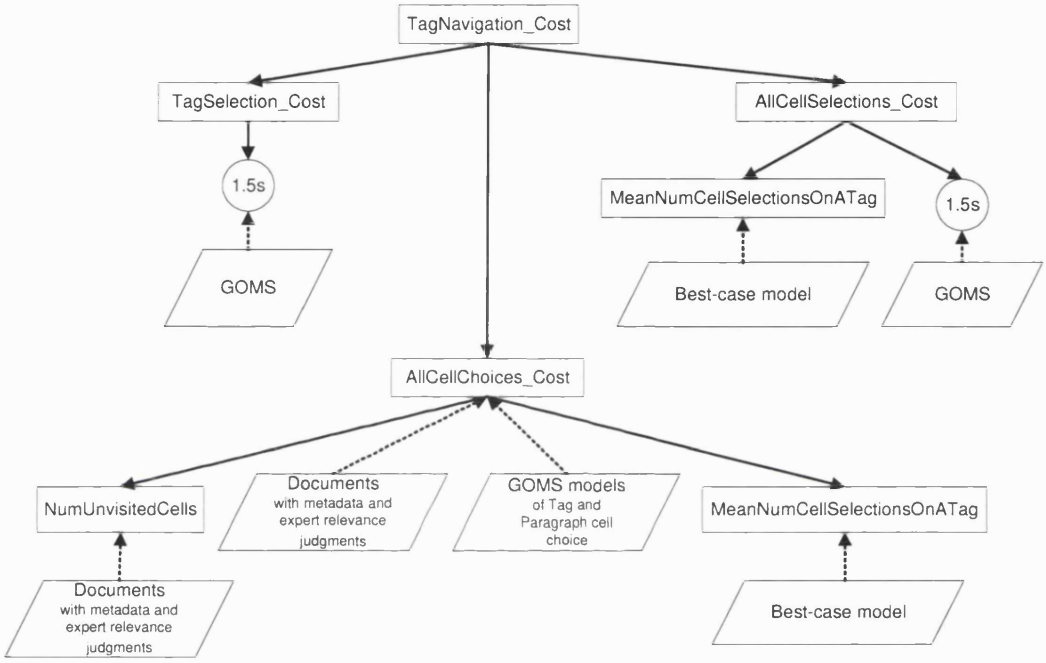


Figure 70 An illustration of how the tag selection cost is broken down into various components, ▭, and what data, ▱, is used to set their values and accompanying coefficients and constants ○

The TagSelection_Cost can be estimated as a constant of 1.5s from the GOMS model (see section 2.2.3.1, and Appendix 4). Similarly AllCellSelections_Cost can be estimated as that same constant multiplied by the number of paragraph cell choices made. Nevertheless, the number of selections will depend on the profitabilities of paragraph cells which are not known when tags are being chosen. However, an expert user could reasonably know (or have a feel for) how many paragraph cells on average

are selected for a tag (e.g. MeanNumCellSelectionsOnATag). Hence the AllCellSelections_Cost is estimated using Equation 7.

AllCellSelections_Cost = 1.5 * MeanNumCellSelectionsOnATag

Equation 7

AllCellChoices_Cost, which is the time taken to search for and decide on each of the paragraph cells selected, is only modelled for the co-occurrence based strategies. For the AppOnly and LtoR strategies the search is carried out in parallel with the selection action and is thus subsumed by the cell selection cost (see Appendix 4). For the co-occurrence strategies the AllCellChoices_Cost can be estimated using the number of unexplored cells on the tag, NumUnvisitedCells, and the MeanNumCellSelectionsOnATag. An expert user could estimate NumUnvisitedCells, by glancing at the row of cells on the tag. As argued above an expert user can be assumed to know MeanNumCellSelectionsOnATag. Thus the estimation of AllCellChoices_Cost could be modelled by identifying the relationship between it and two variables[24]: NumUnvisitedCells and MeanNumCellSelectionsOnATag.

A function to determine AllCellChoices_Cost was therefore produced by statistically fitting a mathematical function to the relationship between AllCellChoices_Cost and NumUnvisitedCells and MeanNumCellSelectionsOnATag. The GOMS models were used to calculate AllCellChoices_Cost for all tags on all documents. For each tag, the cells were selected in order of increasing scent, and the cumulative search time for that tag was recorded as AllCellChoices_Cost. The functions below (Equation 8 and Equation 9) were found to fit the data, using non-linear regression, with an R^2 of 0.996

[24] There are of course many other factors that go into determining the exact value of the AllCellChoices_Cost, but these could not feasibly be estimated by even the most expert of users.

for both the CoocOnly strategy and Cooc&RestrictedApp strategies[25]. Although the accuracy of the estimates resulting from these functions is highly variable (standard deviation of the percentage difference was 30%), they tend to overestimate the actual AllCellChoices_Costs by on average 29%. Some degree of error is inevitable when the estimation is based on the limited information available to the user.

$$\text{AllCellChoices} = -0.002160274 * \text{NumUnvisitedCells}^{1.449154123} * \begin{pmatrix} -275.7538423 - \\ 417.2337886 * \text{MeanNumCellSelectionsOnATag} + \\ 8.289121394 * \text{MeanNumCellSelectionsOnATag}^2 - \\ 0.058220964 * \text{MeanNumCellSelectionsOnATag}^3 \end{pmatrix}$$

Equation 8 AllCellChoices_Costs for the CoocOnly strategy

$$\text{AllCellChoices} = 0.0003974 * \text{NumUnvisitedCells}^{1.4564} * \begin{pmatrix} 1708.3665 + \\ 2334.9787 * \text{MeanNumCellSelectionsOnATag} - \\ 46.9831 * \text{MeanNumCellSelectionsOnATag}^2 + \\ 0.3335 * \text{MeanNumCellSelectionsOnATag}^3 \end{pmatrix}$$

Equation 9 AllCellChoices_Costs for the Cooc&RestrictedApp strategy

Taken together these estimates produced good estimates of navigation costs allowing the model to make foraging decision based on reasonably well founded assessments of the costs implied by different interaction paths. The next chapter briefly confirms the

---

[25] Clearly, the precision in the constants used in these equations is excessive given the degree of error in the predictions they produce. However imposing an arbitrary limit on the number of decimal places might reduce the models' fit (especially given the presence of cubic terms). An alternative approach would be to determine how many decimal places can be removed from each constant without significantly changing the resulting function. It was however deemed more expedient to use the constants without reducing the number of decimal places.

validity of these models by comparing the expected cost of exploiting tags, to the actual costs incurred when they are chosen.

Functions have now been defined for each of the components that make up TagNavigation_Cost in Equation 6. The results of these can therefore be summed, together with the expected reading time estimate (described in the next section), to calculate the expected tag exploitation time. An important feature of this approach is that the only variable which needs to be independently estimated is MeanNumCellSelectionsOnATag. All of the remaining variables have been estimated based on data from the documents (described in section 3.2.1) and the GOMS model. The method used to estimate independent variables such as MeanNumCellSelectionsOnATag will be discussed at the end of this section.

Next the function for estimating SectionNavigation_Cost must be defined. Equation 10 shows a straightforward breakdown, which relates directly to the series of actions that are taken once a section is chosen (see Figure 68). First, it can be noted that the contents of the summation in Equation 10 are identical to Equation 6. Hence Equation 10 can be approximated by Equation 11, where MeanTagNavigation_Cost is the typical navigation cost expected for a tag in the section being investigated. This is a reasonable simplification because an expert user could not be expected to estimate the individual costs of each of the individual tags they might exploit. Thus, the estimation of each of AllTagChoices_Cost, MeanTagNavigation_Cost and MeanNumTagSelectionsOnASection will now be discussed. The way each of these is broken down is illustrated in Figure 71.

$$\text{SectionNavigation_Cost} = \text{AllTagChoices_Cost} + \sum_{\text{for all tag selected}} (\ \text{TagSelection_Cost} + \text{AllCellChoices_Cost} + \text{AllCellSelections_Cost}\ )$$

Equation 10

$$\text{SectionNavigation_Cost} = \text{AllTagChoices_Cost} + \text{MeanTagNavigation_Cost} * \text{MeanNumTagSelectionsOnASection}$$

Equation 11

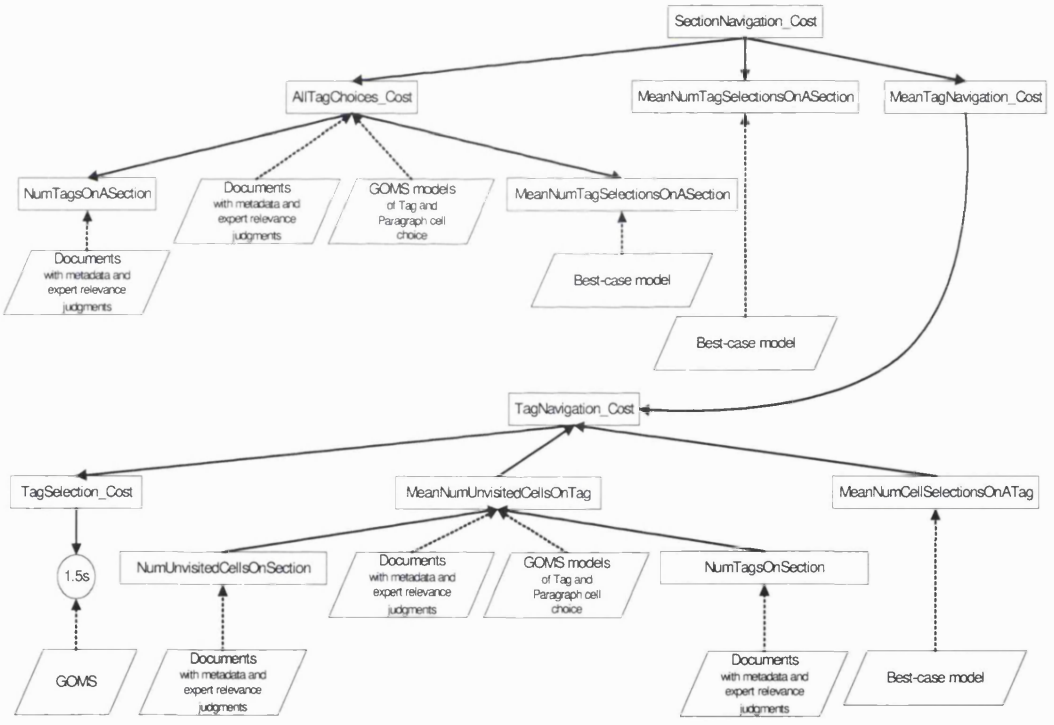


Figure 71 An illustration of how the estimation section selection cost is broken down into various components, ▭, and what data, ▱, is used to set their values and accompanying coefficients and constants ○

The AllTagChoices_Cost refers to the time taken to search for and decide on each of the tags that are pursued for the section being exploited. As with AllCellSelections_Cost there are many factors that determine how much time this will take, but most of them could not feasibly be employed by the user in assessing exploitation cost. Two variables that are available are the number of tags on the section, NumTagsOnSection, and the number of tags likely to be selected, MeanNumTagSelectionsOnASection. NumTagsOnSection can be visually estimated by glancing at the metadata tree, and MeanNumTagSelectionsOnASection is assumed to be part of the expert user's expertise.

The estimation of AllTagChoices_Cost was modelled by statistically fitting a mathematical function to the relationship between AllTagChoices_Cost, MeanNumTagSelectionsOnASection, and NumTagsOnSection. The GOMS model was use to calculate AllTagChoices_Cost for every section in every document. For each section the tags were selected one at a time, and the cumulative search time on that section was recorded as AllTagChoices_Cost. The function in Equation 12 was found to fit the resulting data, using non-linear regression, with an R^2 of 0.988. The accuracy of this estimate was high, with the mean percentage difference being –2%, although rather variable (standard deviation of the percentage difference was 17%).

$$\text{AllTagChoices_Cost} = \left(3.200879425 * \text{NumTagsOnSection}^{0.975436633}\right) * (2.350252505 + 1.450372842 * \text{MeanNumTagSelectionsOnASection})$$

Equation 12

The next variable to be estimated in Equation 11 is MeanTagNavigation_Cost. This refers to the average amount of time spent on selecting the tag, searching for relevant cells and selecting them. This can be estimated using the function for estimating

TagNavigation_cost described above outlined in Equation 6. Instead of using Equation 6 to estimate the navigation cost of exploiting individual tags, it can be use to estimate the exploitation time for the *average* tag in the section being considered. The calculation of all the components in Equation 6 requires two parameters: MeanNumUnvisitedCellsOnTag and MeanNumCellSelectionsOnATag.

Firstly, MeanNumUnvisitedCellsOnTag is required to calculate the AllCellChoices_Cost parameter in Equation 6. It can be estimated from the number of unvisited cells on the section tag and the number of leaf tags in the section. These quantities are accessible to the user, requiring a glance at a section tag's row in the central grid to roughly assess the number unvisited cells, and another glance below the section tag to roughly assess the number of leaf tags. A function relating these two parameters to the actual mean number of unvisited cells on the leaf tags in the section was statistically derived from the metadata on the documents. A linear regression of the number of tags and number of unvisited cells on a section produced a function which predicted the mean number of unvisited cells on a section tag (see Equation 13) with a low but acceptable fit ($R^2 = 0.276$). The accuracy of this estimate was not high, with the mean and standard deviation of the percentage difference between actual and predicted being 23%, and 58% respectively. However Figure 72 shows that the error tends to 0 when the actual number of tags is at its mean of 3, but increases positively (i.e. overestimates) as the number of tags decreases and vice versa. This is to be expected from a linear regression, and this is the best that can be done, since the data did not merit the application of any non-linear functions. Additionally, given the limited information being used to make this estimate, the 23% mean percentage error is very reasonable.

MeanNumUnvisitedCellsOnTag = 2.412 + NumUnvisitedCellsOnSection * 0.152 – NumTagsOnSection*.298

Equation 13

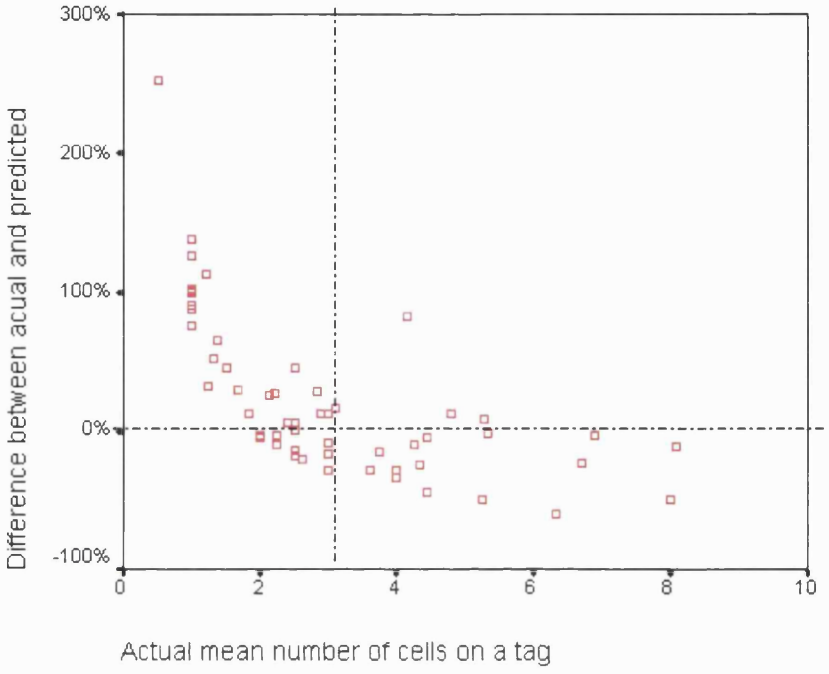


Figure 72 The percentage difference between the predicted and 'actual mean number of unvisited paragraph cells' (shown as a percentage of the actual number of unvisited cells). The cross hair drawn with a dotted line, illustrates that the error tends to 0 when 'actual mean number of cells on a tag' tends to, 3, its mean.

Finally, the estimation of the two key parameters, MeanNumCellSelectionsOnATag and MeanNumTagSelectionsOnASection, must be discussed. These are the only parameters in the calculation of expected navigation time which are not derived from the documents and GOMS models. Since these depend on the behaviour of the model, i.e. how many selections are actually made, they are difficult to define analytically. A simple approach however is to arrive at them numerically, by running the model with these parameters set to arbitrary seed values (i.e. 1). Once the model has run, the actual values for MeanNumCellSelectionsOnATag and MeanNumTagSelectionsOnASection

are calculated from the results. These values are used as inputs and the model is run again, the values are recalculated, and compared with those obtained on the previous run. If they are different, then the model is run again, and again, until the parameters settle.

This approach has several dangers, the final values of the parameters may be dependent on the initial seed values, and the parameters may never settle. Fortunately, the final values of the parameters were not in fact dependent on the initial parameters. However, in some cases the parameters did not settle on a particular value but oscillated between two or three. On these occasions mean values were used as the final value for the parameters.

2.2.3.2.2 Estimating paragraph handling time

The paragraph handling time needs to be estimated in order to allow the expected cost of exploiting paragraph cells to be computed. The expected cost of exploiting paragraph cells is needed in turn to compute which cell are to be chosen.

Fortunately the paragraph handling model (in section 2.2.3.1) can be used to make this estimate. The paragraph handling model defines how the *actual* paragraph handling time is calculated using the paragraph's actual information value and length. The paragraph handling model can be employed because GridVis gives the expert user what they need to estimate these two parameters: the information scent suggests the amount of information there is, and the width of paragraph cells indicates how much text there is. Hence the *expected* paragraph time can be calculated in a similar way to the *actual* times.

However, when estimating the paragraph handling time of an unseen paragraph there is uncertainty about whether or not the paragraphs will be relevant, and thus read in depth. This poses a difficulty since in Equation 5, which is repeated in Equation 14 and

Equation 15, different functions are needed for relevant and irrelevant paragraphs. The approach taken was to estimate the proportion of paragraphs that are relevant for cells of a given scent. The paragraph handling cost is then calculated using each of the equations for relevant and irrelevant paragraphs; the result of each is weighted by the likelihood of the paragraph being relevant or irrelevant respectively, these weighted costs are then summed (see Equation 16).

$$\text{ParaTime}_{\text{irrel}}(\text{NumWords}, \text{Relevance}) = \text{NumWords} * 0.015045 + \text{Relevance} * 2.128401 + 1.239530$$

Equation 14 the time taken for irrelevant paragraphs

$$\text{ParaTime}_{\text{rel}}(\text{NumWords}, \text{Relevance}) = \text{NumWords} * 0.116919 + \text{Relevance} * 0.951996 + 4.483907$$

Equation 15 the time taken for relevant paragraphs

$$\text{Expected Paragraph Time} = \text{Prob}_{rel}(\text{Relevance}) * \text{ParaTime}_{\text{rel}}(\text{Relevance}, \text{NumWords}) + \text{Prob}_{irrel}(\text{Relevance}) * \text{ParaTime}_{\text{irrel}}(\text{Relevance}, \text{NumWords})$$

Equation 16

This begs the questions of how to estimate Relevance, NumWords, Prob_{rel} and Prob_{irrel}. The rest of this section will explain how these estimates were arrived at.

Relevance can simply be estimated by information scent, since the scent is the expected information value, which is measured in units of relevance.

Paragraph size, NumWords, can be estimated from the width of the paragraph cell. An expert user of GridVis is assumed to have a feel for the relation of cell width to paragraph size which is rapidly calibrated to the particular document being used. However the performance of adult humans in discriminating between different widths of light and dark bars rapidly decreases below about ten cycles per degree (Ware 2000).

Since with a standard monitor at typical viewing distances there are forty pixels per degree (Ware 2000) this corresponds to about four pixels. Estimates of paragraph size are therefore modelled by calculating, for each document, the number of words that correspond to four pixels on the grid. This is then multiplied by the number of four pixel widths that fit into the cell in question. This method of estimating paragraph widths results in a mean error of about 13%.

To obtain the $Prob_{rel}$ and $Prob_{irrel}$ functions it was necessary to examine the relationship between scent and paragraph relevance. For each paragraph cell choice strategy the cell scent and paragraph information value were computed, for all instances on all tags. The cumulative proportion of relevant paragraphs as scent increases was then calculated and plotted for each strategy. The resulting curves were then fit using various functions; a logarithmic function was used for the AppOnly strategy, and cubic curves were used for the others[26]. The resulting equations predict the probability that a paragraph will be relevant given the information scent of a corresponding paragraph cell. These equations are listed in Table 23. When the scent is either very high, or very low these equations can produce a result outside of the 0 to 1 range that is meaningful for a probability. On these occasions a cut off limits these 'extreme' results to 0 or 1 respectively.

Strategy	Equation	F	R^2
LtoR	0.076 + 0.9701 * scent - 0.4551 * scent^2 + 0.0775 * scent^3	33011.7	0.975
AppOnly	0.6975 + 0.3292 * Log(scent)	56647.1	0.957
CoocOnly	-0.0153 + 0.002 * scent + 0.2967 * scent^2 - 0.0625 * scent^3	121463	0.993
CoocRestrictedApp	-0.0178 - 0.0029 * scent + 0.3057 * scent^2 - 0.0619 * scent^3	136380	0.994

Table 23 Equations for estimating the probability of a paragraph being relevant given the scent of a paragraph cell

[26] This was done using SPSS's curve estimation regression procedure.

2.2.3.2.3 *Estimating the mean patch choice time*

The mean patch choice time, t_b, is the mean time it takes to choose the next patch (i.e. section tags, leaf tags, instances) using GridVis. It is needed in Equation 3 of the optimal diet selection algorithm in order to determine whether to select another item.

A separate value of can t_b was calculated for section tags, leaf tags and paragraph cell[27]. t_b was estimated in much the same way as MeanNumCellSelectionsOnATag and MeanNumTagSelectionsOnASection; it is taken as the mean value of the search times generated by the best-case model. These search times are the product of running the best-case model to calculate the *actual* cost of searching for tags. Means of these search times are taken for sections, tags, and instances.

There is again a problem of circularity here in that the value of t_b helps determine which selection are actually made, which in turn determines the resulting set of search times, from which t_b is calculated. This problem was tackled together with that for the number of selection estimations, by setting t_b to 1 and iteratively rerunning the model, feeding resulting values for t_b in, on each pass, until the values stabilised. The values for t_b did not depend on the what seed value was used, and almost always settled. On a few occasions, the t_b value for paragraph cell choices oscillated between two or three values. In these cases the mean value was used as discussed above 'Estimating GridVis navigation time'.

2.3 Section summary – an overview of the best-case model

The best-case model is first defined at the strategy level then at the implementable performance level. At the strategy level, the model uses the optimisation rule, which is

---

[27] The b in $t_{b_{itemType}}$ is used because the search is in fact a between-patches search.

formalised using information foraging theory's optimal diet selection algorithm, to select the most profitable sections, tags and paragraph cells. Strategies for efficiently gathering scent for tags and paragraph cells are then defined.

At the performance level the model is defined as scent and cost models. The scent models, allow the simulation of human gathering of scent for making decisions between items; they attempt to capture the expert's knowledge of the relationship between cues in the interface and information value. For tags, scent is mapped directly to the actual amount of information available, this is called perfect scent. For paragraph cells, separate models are constructed for each cell selection strategy. These rely on the cues present in the GridVis display to estimate the amount information available. The cost models are of two types, models of actual cost and models of expected cost. The models of actual cost are used to calculate the task time. There are NGOMSL models of tag and paragraph cell choice, and a statistical model of reading time which is a fit to data from the paper performance experiment (described in section 3). The models of expected cost are needed to complement scent information in deciding between items. These models attempt to mimic expert's experience with GridVis. The estimation of expected GridVis navigation time is modelled by breaking the navigation time down into simple components and fitting functions to these using parameters and data from the NGOMSL models and documents. The expected paragraph handling time is modelled by combining the model of actual paragraph handling time with data from the NGOMSL models and documents. Finally, the average patch choice times need to be estimated for use by the optimal diet selection algorithm in computing the choice of items. This, along with two parameters needed for the model of navigation time, is estimated numerically from the results of iteratively running the best-case model.

3 Paper performance experiment

This experiment was designed to obtain two types of data needed by the best-case analysis. Primarily, it measures the performance of expert users in a goal-directed search task. This is the benchmark performance against which GridVis must compete. An additional aim was to get data on the times taken 'to reject a paragraph as irrelevant' and 'to accept a paragraph as relevant'. This data is used to estimate some of the parameters in the cost model (see chapter 6, section 2.2.3).

Twenty four Sainsbury's head office employees were given a goal-directed search task; they were asked to look for the text relevant to written queries. Each participant completes this task for six queries, which varied in difficulty. For hard queries the headings of relevant paragraphs included one of the concepts used in the query, whereas for easy queries two were included. One 'hard, and one 'easy' query was set for each of three documents. These three documents, were varied in length: one was short (5pgs), one medium (8pgs), and the final one long (14pgs). Document length and information were varied since it was hypothesised that increased length and information need difficulty would impair performance with paper, but not with GridVis. For each document, the subjects were asked to highlight the relevant text and track their progress down the page with the tip of a pen. This was captured on video so that the time spent one each paragraph could be measured.

3.1 Sample

The twenty four subjects used for this experiment were all from Sainsbury's head office logistics department. These are numerate knowledge workers who manage and develop the supermarket's supply chain. They can be considered expert document users since they have gone through higher education and have spent their working lives dealing

with documents. Working at Sainsbury's has given them a rich knowledge of the supermarket industry and experience of the kind of documents being used in the experiment. Importantly, they 'share the corporate framework and values [...] they necessarily share collective knowledge structures, and will experience information needs in similar ways' (Bryce 1997). They are also familiar with the interests and priorities of the industry. In order to ensure that this was indeed the case participants were only included if they had been with Sainsbury's for more than six month, and did not work as external consultants.

As discussed in chapter 5, section 4, recruiting participants from a corporate context is difficult when a large amount of time is required by the experiment. For this reason, in order to recruit 24 suitable subjects, it was deemed necessary to keep the time needed from each participant down to around half an hour.

3.2 Materials

A set of three documents, with two information needs each, were used in the paper performance experiments and with the best-case model. These documents and queries were carefully constructed to examine the effect of document length and information need difficulty, while minimising the effect of extraneous factors. Each paragraph was then rated for relevance to the queries by a set of expert judges.

3.2.1 The documents

This section describes how the documents were constructed. It discusses the choice of document type and the modifications used to produce variations in document length and information need difficulty. An appropriate document type was selected from among those in use at Sainsbury's. The document type used was a weekly news-digest which is an internally produced compilation of short articles from the national and trade press

that relate to both Sainsbury's and retail logistics, the particular area in which the participants and frontier users work. The documents used were artificially constructed to yield differences in document length and query difficulty whilst conforming to the news-digest document type in terms of content and format. The choice of document and the way in which their lengths were varied are discussed below, further considerations involved in the construction of these documents are covered in Appendix 7.

3.2.1.1 Choice of document type

A number of factors determined the choice of document type. The documents needed to be relatively easily understood by Sainsbury's staff. This would avoid two problems encountered in the experimental evaluation (see chapter 5 section 4); there the difficult nature of the documents made the experimental task unnecessarily hard and consequently placed impractical restrictions on the number of suitable subjects available. The 'industry updates' are written in readable style, accessible to anyone familiar with workings of the supermarket industry.

A related advantage of this document type is that its more pedestrian content is more easily described using metadata. Whereas the statistical analysis documents used in the empirical evaluation put across fine and original points for which no single word or phrase was an adequate signpost (see chapter 5 section 3.4), the industry updates deal with the common events and objects of retail business for which powerful labels already exist.

Finally, the industry update's modular format allowed documents to be constructed to fit the experimental tasks. These documents are essentially made up of a set of news clippings organised into topics. Articles can therefore be collected together to produce a

series of industry updates which vary along the dimensions being examined in the evaluation (specifically document length and information need difficulty).

However, one important problem with this choice of document type is that they are not generally read in a goal-directed manner. The information needs interviews (chapter 4 section 2.2) revealed that this document type tends to be read for interest; depending on the time available people will, on a weekly basis, read or skim through the entire document looking for news of interest. Nevertheless, although this document type is not currently used in a goal-directed manner perhaps it might be if it that became a less time consuming affair. There are many possible motivations for goal-directed search within news digest documents. For example, a frontier user noted that he often wanted get a better understanding of the history of events by referring back to related stories in previous industry updates. However if these motivations are not high priorities, doing the search is not worthwhile. Hence, although the document type being used in this investigation is not typically read in a goal-directed manner, it could well be if the cost of doing so was low enough.

3.2.1.2 Varying document length

The modular nature of the selected document type allowed the length of the documents to be varied by changing the number of news clippings included from past 'industry updates'.

The document lengths were chosen to get the maximum range of document size while satisfying several constraints. The most important of these constraints was to keep the time needed for each trial of the paper performance experiment (described in section 3.1) to 30 minutes or less. Making the experiment much longer than this makes recruiting participants from the workplace very difficult. It was also important that the

documents remained recognisable members of the document type selected. The documents needed to be long enough to have reasonable quantities of non-relevant 'distracter' material in subsections containing relevant paragraphs, while including each of the document type's other sections. Relevant paragraphs also needed to be located in similar contexts across documents. In the longer documents a section containing relevant paragraphs would also contain distracters. However if the documents got too short then it was not possible to add distracter paragraphs to sections containing relevant paragraphs, without making the section disproportionately long. The final lengths chosen were 5pgs, 8pgs and 14pgs (see Appendix 7 for more detail on this).

3.2.1.3 Varying information need difficulty

In the evaluation of Superbook (Egan et al. 1989) the difficulty of information needs was varied. Easy information needs had answers in sections whose titles contained words matching query words, difficult ones did not. A similar technique was used here. However, in order to minimise distortion of the document type and the original material this procedure had to be adapted as follows.

Each query essentially imposed two constraints on what constitutes relevant material (see Figure 73).

What is Tesco doing to keep ahead in home shopping?

Figure 73 A query with two constraints on relevance, indicated by the grey panels.

Information needs were made easy by including words relating to both of the constraints in the headings of relevant paragraphs. They were made 'hard' by only including keywords relating to one of the constraints in the headings of relevant paragraphs.

Distracter paragraphs and headings were added to prevent the task from boiling down to simple keyword matching alone. These were non-relevant paragraphs which sported headings relating to one of the query constraints. Half of the distracters related to each constraint. The number of distracters was roughly doubled from short to medium, and again from medium to long.

3.2.2 Developing representative queries

A set of queries were developed to provide the basis for the goal-directed search tasks that GridVis was being evaluated upon. These were developed such that they were representative of the kinds of information needs that people at Sainsbury's have when reading 'industry updates'. The queries were derived from the data obtained during the information needs interviews described in chapter 4 section 2.2, which were conducted to produce the metadata and the metadata taxonomy.

In these interviews the interviewees were asked to go though an 'industry update' and decide which of the articles they would have read at any length. They were then asked to try to state what information they were interested in acquiring by reading the articles that they selected.

This resulted in a pool of crudely stated information needs from which suitable queries were selected and refined. The queries were selected on the basis that they contained two constraints on what material counted as relevant i.e.:

Are we looking at selling the JS's interests in the US?

As opposed to:

What are the advertisements going to be about?

They were then refined, in order to remove the ambiguity inherent in the interviewees initial verbal statements e.g. from, "What's going on in the other JS businesses?", to "What is going on in other JS businesses within the JS group?".

3.2.3 Relevance judgements

Three Sainsbury's frontier users acted as judges. They were given the three documents and six queries (two for each document). For each query, they were asked to read the associated document thoroughly, and rate each paragraph on the 0 to 3 relevance scale used in the empirical evaluation (see chapter 5 section 3.1.4). Their ratings were aggregated using the procedure described in Appendix 1.

3.3 Method

The experiments were carried out in a standard work setting in the Sainsbury's head office. This office, typical of corporate office space, consisted of a large open plan room with about eighty desk spaces separated by waist height partitions. The office is not very noisy, but there are typically one or two audible conversations going on at any time. This venue was chosen because it was both practical and representative of workplace environments. One desk was used for the experiment; a digital video camera was set up on the desk using a tripod so that the documents were always in view. The experimental task was designed to gather the data required while remaining as similar as possible to actual workplace reading.

The main challenge was to record the amount of time spent on each paragraph without substantially intruding on the task. This challenge was addressed by asking each reader to track their progress as they read the document with the tip of a pen. This is minimally intrusive since using some kind of pointer to focus reading is a widely practiced technique, using a pointer in this way feels natural and unencumbering. The participants

were video recorded, while they read the document using the pen, so that the time spend on each paragraph could be measured.

A second challenge was to make sure users performed the goal-directed reading task in earnest. This might well have been a problem in the experimental evaluation reported in chapter 5 section 3. There seemed to be a tendency for the subjects to look for relevant information while stopping short of thoroughly confirming its relevance by actually processing the information. If so the task was subtly shifted from that of 'goal-directed reading', an important workplace task, to something like 'goal-directed hoarding'; a slightly different and quite artificial task, where information that *might* be relevant is gathered for latter perusal. Hence the task had to be set such that it required the user to do goal-directed search for actually, rather then possibly, relevant material. To do this each subject was asked to selectively highlight the relevant words/sentences/phrases. They were also asked to imagine that they needed the information for an impending meeting. The selective highlighting requirement forced the reader to go through relevant text; it was a request both familiar to the subjects and appropriate to the task setting. It also provided a means of measuring paragraph reading times to give estimates of the times taken 'to reject a paragraph as irrelevant' and 'to accept a paragraph as relevant'. The imaginary scenario gave the subjects a motivation for the task that served to encourage goal-directed search. This approach did however have the drawback that it increased the time taken 'to accept a paragraph as relevant'. However, highlighting during reading is common practice (Schilit, Price & Golovchinsky 1998) and thus is a valid type of reading to study. Also estimates of the task times that would be achieved without highlighting could be calculated using the data from the data gathered from this experiment and existing data on reading times (Card, Moran & Newell 1983).

The document presentation order was chosen so as to deal with two learning effects. Since there were two questions per document the subject might find it a little easier to answer the second question on a document. To minimise this effect the document presentation order was chosen to maximise the separation between queries on the same document. This should have allowed learning to fade by increasing time to a document's second presentation and by supplying material that interfered with what was learnt. Another potential problem is that people tend to improve at a task as they do it. To deal with this the presentation order was counterbalanced for length and question difficulty (see Table 24).

Order A (12 subjects)		Order B (12 subjects)	
Long document	Q 1 hard	Long document	Q 4 easy
Medium document	Q 2 easy	Medium document	Q 5 hard
Easy document	Q 3 easy	Easy document	Q 6 hard
Long document	Q 4 easy	Long document	Q 1 hard
Medium document	Q 5 hard	Medium document	Q 2 easy
Easy document	Q 6 hard	Easy document	Q 3 easy

Table 24 The document and query presentation orders

The experiment was conducted using a fixed procedure comprising of instruction, a training task, an experimental task, and a post experiment interview. The instruction was begun by asking each subject to read a half page description of what needed to be done in the experiment. The experimenter then talked through this, thereby physically illustrating the instructions and providing an opportunity to clear up any misunderstandings. Each subject was then asked to go ahead with the task using a short, three page, 'practice' document and query. This training task made it possible to clear up misunderstandings before the experimental task. For the experimental task each subject was presented with a pile of six documents, each one topped by a separate sheet with a query on it. Each subject was asked to go through these in order. First they read the query. Since there is a tendency to forget the query, it was then placed in view for

future reference. Then they went through the document as they saw fit. Once they had finished with a document, they went straight on to the next one. Once all six documents were finished, a short post experiment interview was conducted which attempted to uncover the subject's reading strategies.

3.3.1 Data transcription

The video footage was used to gather data on reading times and retrieval performance. Transcription of this data allowed the retrieval metrics detailed in section 4 to be computed. It also allowed the calculation of estimates for the times taken 'to reject a paragraph as irrelevant' and 'to accept a paragraph as relevant' which are needed as input parameters for the best-case model (see section 2.2.3.1.3 on modelling actual paragraph handling time).

The time spent on individual paragraphs was measured using video footage of each reader's pen movement and page turns. The procedure for doing this is described in detail in Appendix 5.

Each paragraph was categorised according to the type of reading strategy, the time measurement's reliability and its position on the page. The most important distinction being whether the paragraph was judged relevant (and therefore read in detail and highlighted) or rejected as irrelevant. This categorisation allowed performance measures such as recall and precision to be computed as well enabling separate mean reading times for paragraph acceptance and rejection to be calculated. The remaining categorisations were all introduced to reduce the variance (and improve the quality) of the paragraph time data. These were however not used in the final analysis since they were not effective in reducing variance.

The full data transcription described here was carried out for sixteen of the twenty-four subjects. Sixteen subjects provided enough data to give reliable estimates of paragraph reading times, search precision, 'proportion of time spent reading' and 'proportion of reading time with relevant'. The task time and paragraph selection data was calculated over all twenty four subjects.

4 Performance Metrics

A number of performance metrics were calculated from the data obtained via the best-case model and the paper performance experiment; these included the traditional IR metrics of precision and recall, as well a number of additional metrics devised to capture the performance tradeoffs between paper and GridVis.

Traditional and interactive recall and precision measures were calculated as detailed in chapter 2 section 5.1.2.1. A refinement of the recall measures was also calculated which split the measure into two, Core Recall – the recall of highly relevant information (paragraphs with a relevance rating of 3), and Peripheral Recall – the recall of somewhat relevant information (paragraphs with a relevance rating of 2 or 1). This distinction was made since in the empirical evaluation reported in chapter 5, section 3, it had become clear that paragraphs were not of equal significance; some were vital, and some were nice to have. In goal-directed search this is an important distinction. GridVis would be a valuable tool if it rapidly delivered all the vital paragraphs, at the cost of missing those that were just nice to have. Hence to capture this possible trade off Core Recall, and Peripheral Recall were calculated as follows:

$$\text{Core Recall} = \frac{\text{No of paragraphs of rating 3 deemed relevant}}{\text{Total number of paragraph of rating 3}}$$

$$\text{Peripheral Recall} = \frac{\text{No of paragraphs of rating 1 or 2 deemed relevant}}{\text{Total number of paragraph of rating 1 or 2}}$$

The total task time was calculated for each query. This was measured as the time from the moment the document is in position in front of the reader, to the moment they begin to turn the last page.

An additional metric was developed to complement interactive precision in measuring the cost of dealing with irrelevant material. This metric is the 'proportion of reading time spent with relevant'. It is needed since, unlike interactive precision, it addresses cost more directly. It is also sensitive to benefits of the strategies used in reading paper documents. With paper, the reading strategies adopted tended to involve quickly scanning irrelevant material and iteratively inspecting relevant material, taking more time with each iteration. This issue is revisited in chapter 7 section 1. Thus by assigning equal weight to paragraphs read carefully, to those that were skimmed, interactive precision exaggerates the problem of dealing with irrelevant material. The 'proportion of reading time spent with relevant' paragraphs measure was calculated as:

$$\text{Proportion of reading time spent with relevant} = \frac{\text{Time spent on relevant paragraphs}}{\text{Total time spent on all paragraphs}}$$

Finally, a metric was developed to measure the cost of navigation. This measure is needed because there is trade-off in the use of GridVis between ability to avoid irrelevant material and the investment in navigation required. The measure, 'proportion of task time spent reading' reveals the amount of time devoted to navigation. This was calculated as:

$$\text{Proportion of task time spent reading} = 1 - \frac{\text{(Time spent on document)-(Time spent on all paragraphs)}}{\text{Total time spent on document}}$$

5 Chapter summary

This chapter presents a new evaluation method: 'best-case analysis'. A best-case analysis is intended to allow the evaluator to robustly and quantitatively assess the potential utility of a design or class of designs for interactive software without the need for large numbers of participants. It essentially compares the best performance that can be expected from a human operator of the design in question with that of a benchmark, which is the most prevalent or effective competing tool.

In this chapter the best-case analysis method is described and then the detail of its' application to GridVis is presented in four parts. First, the construction of the best-case model of GridVis use is described. This is initially described in terms of the strategy for using GridVis, and then in terms of the implementation of this strategy used in the model. Both in the strategy adopted and the details of the implementation, the model attempts to achieve the highest performance while respecting the limits of human cognition and making reasonable assumptions about the experience of expert users. The second part describes the materials, namely, the documents and queries, used both by the model and by human participants. The third describes the paper performance experiment, in which data for use in setting parameters of the best-case model are collected and the performance of the benchmark, namely, paper, is measured. The fourth and final part reports the performance metrics calculated from the data collected via the best-case model and the paper performance experiment.

Chapter 7 Results from the best-case analysis

Chapter 6 described the best-case analysis of GridVis which involved an experiment on peoples performance with paper and the construction of the best-case model of GridVis. The best-case model of GridVis was run to yield a number of performance metrics for comparison with those obtained from the paper performance experiment. This chapter examines the results from the paper performance experiment. These shed light on the strategies used with paper which suggest that paper performance can be validly compared to the best-case model of GridVis use. The results for the best-case model of GridVis are then compared to the performance with paper to reveal trade-offs. The results also reveal that performance with GridVis depends on the paragraph selection strategy used.

1 The benchmark – performance and strategies with paper

This section reports the results from the paper performance experiment described in chapter 6, section 3. The data gathered is first used to reveal the kinds of high-level reading strategies that are used and then overall retrieval performances are reported.

1.1 Reading strategies

Participants in the paper performance experiment employed a wide range of strategies. The data on these were collected using the debriefing interviews and backed up with observational notes taken by the experimenter. The data were initially examined by making detailed notes from recordings of the debrief interviews.

Examination of this data allowed the identification of an important high-level feature of the reading strategies used. All the participants spent time attending to every page in each document. Twenty of the twenty four participants made no use of the table of

contents nor of the document's organisation into high level sections (e.g. retail articles, distribution articles etc). These participants progressed methodically through the document looking at every page albeit spending much less time on pages with no relevant articles. The few who made use of document structure, would either use a section's relevance to determine the reading strategy adopted, or turn directly to the page on which a relevant section began. This suggests that the participants took a similar approach to that embodied by the best-case strategy. The best-case strategy uses an 'optimising rule' where all options are examined, rather than a 'satisficing rule' in which options are examined until a good enough item is found (see chapter 6, section 2.1). The overwhelming tendency of participants using paper to proceed linearly though the entire document suggests that they are not satisficing, but attempting to comprehensively review the document and extract all the information that is relevant. Thus there is an important correspondence between the high-level strategies followed by the best-case model of GridVis and the expert users of paper. This suggests that the results from the best-case and paper can be validly compared.

Initial examination of the data suggested that reading strategies were deployed in a pattern that conforms to the predictions of information foraging theory. The data suggested that although the detail of individual reading strategies varied substantially across participants, their structure remained consistent. Low cost and low fidelity strategies, such as scanning the article titles for relevant keywords, were used first as coarse filters identifying material worthy of further examination. Higher cost, higher fidelity strategies, such as reading the first two lines of a paragraph were then progressively deployed for the material that had passed through the preceding filters. As a result, high cost reading strategies were only used with very relevant looking material; valuable time and effort is not wasted on material which is likely to be irrelevant. This

configuration of reading strategies conforms to qualitative predictions of information foraging theory. Essentially, by holding back the most costly strategies for the most relevant paragraphs, expected cost is being balanced against expected information gain; this has the effect of maximising the rate of information gain.

This conclusion is further evidence that the results from the best-case model can be directly compared to the performance with paper. The decisions made by the best-case model are determined by information foraging theory which assumes that behaviour tends to maximise the rate of information gain. If strategies used with paper conform to the qualitative predictions of information foraging theory, then they too tend to maximise the rate of information gain. This therefore additionally supports the claim that the best-case model and the paper experiment are comparable.

Thus, in order to confirm this initial characterisation of the reading strategies the data was further analysed in a systematic manner as follows. Each strategy or type of reading mentioned or observed by the participants was listed. These were grouped into to a set of strategies, (e.g., scan titles, read titles, scan paragraph fast). These were drawn as nodes on a pathless graph. Every transition between reading types explicitly revealed in the data was marked as a path between nodes or as a tally on an existing path. Separate paths were drawn for transitions made upon positive, negative evidence and for those made regardless of the evidence. The result of this collation of the data can been seen in Table 25 and Figure 74.

		Low cost strategies →									High cost strategies
	From > / To v	Read table of contents	Go to relevant section	Go to non-relevant section	Scan titles	Read titles	Scan paragraph - fast	Read start of paragraph	Scan paragraph - medium	Revisit paragraph later	Read paragraph
Low cost strategies	Read table of contents										
	Go to relevant section	r3									
	Go to non-relevant section		r1								
	Scan titles		r2	r2							
	Read titles		r2		n1						
	Scan paragraph - fast				n3	r5		n1			
	Read start of paragraph				r4, 1						
	Scan paragraph - medium				n2, r6, 1		r1	r3, n1			
	Revisit paragraph later							r2			
High cost strategies	Read paragraph				r5	r1	r1	r2	r7		
	Reject				n6	n4	n5	n6	n7		

*Table 25 All the transitions between reading strategies. The numbers are coded, **r#**, for transitions upon positive evidence (i.e. of relevance), n#, on negative evidence (i.e. of irrelevance) and, #, if transition is made regardless of evidence. This table shows that low cost strategies lead to higher cost strategies which themselves lead on to higher cost strategies until the paragraph is either rejected or read.*

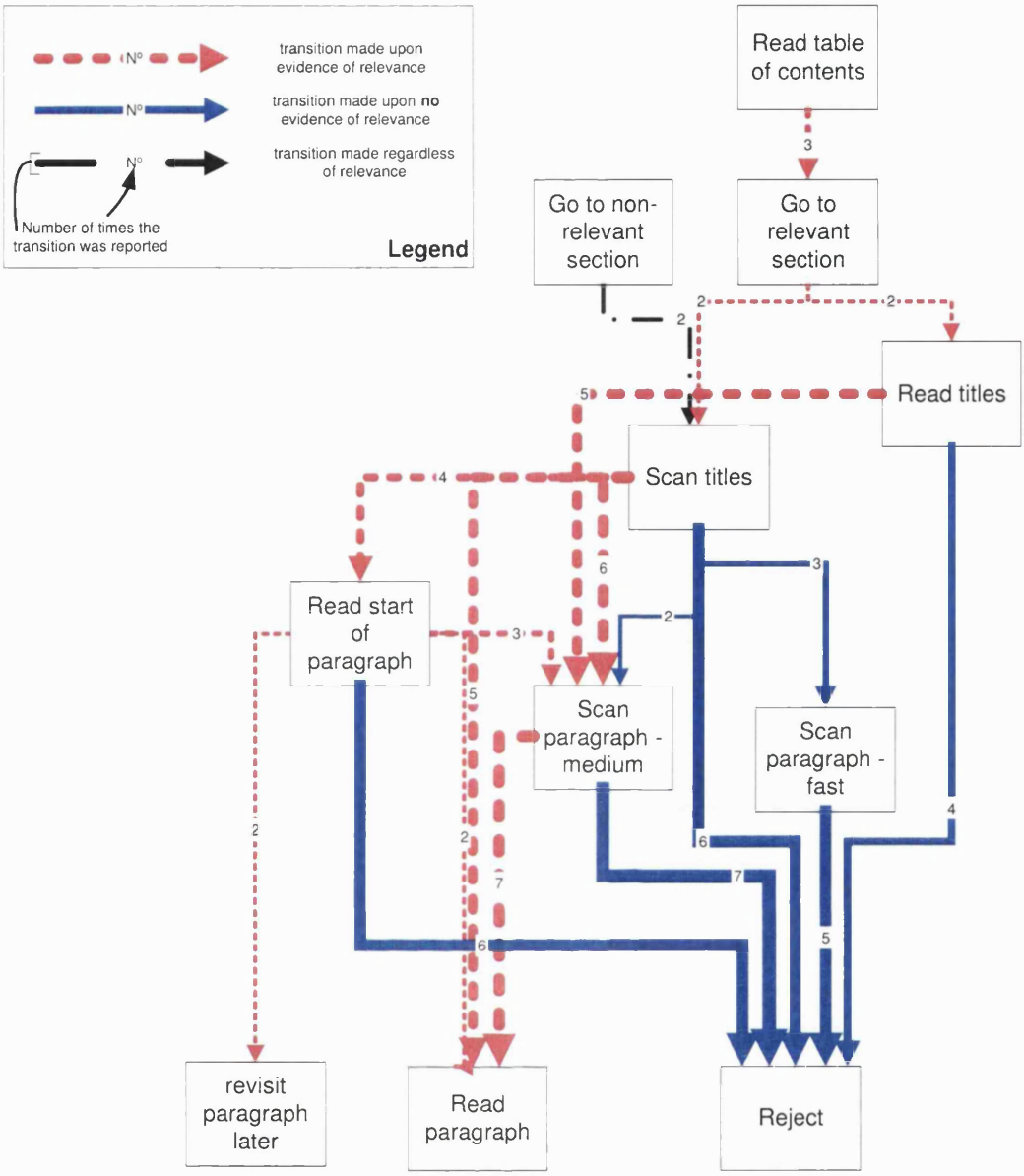


Figure 74 The collated data on reading strategies and the transitions between them, from interview and observational data. Transitions that were only evidenced by one subject have been excluded to clarify the overall picture. By starting at the top and tracing a path downwards, one can appreciate the use of multiple layers of reading strategy, which increase in cost as the expected relevance of material increases.

Figure 74 shows both the consistency in the character of participants' reading strategy and the variety in its details. Starting at the top of the diagram one can trace several paths through the task, each of which conforms to the notion that strategies of higher cost and fidelity are successively applied to relevant-looking material. For example,

some participants started by *scanning the titles*. If a title seemed relevant, they would *read the start of paragraph* before reading the whole thing through or *rejecting* it and moving on to the next section. Otherwise they would double check on its irrelevance by *rapidly scanning* before ultimately *rejecting* it. The subjects varied in the type and number of different strategies (or filters) they invoked. For example, some would only use *read titles* to decide whether a paragraph should be *read* or *rejected*, while others would use several layers of filtering strategy before reading a paragraph in full. These configurations of strategies all show a structure which tends towards minimising cost and maximising information gain. This very same structure was reported by Pirolli and Card in two studies that examined the information seeking behaviour of industry analysts and graduate students (Pirolli & Card 1999). They note how this structure conforms to the qualitative predictions of information foraging theory (Pirolli & Card 1999).

There are some apparent inconsistencies in Figure 74; these are accounted for by the incomplete nature of the data, the removal of infrequent transitions, and variation in the configuration of user strategies. The number of transition flowing into a node often does not equal the number leaving it. This is partially because the data were often incomplete since participants did not explicitly reveal every aspect of their reading strategies. In addition, transitions only reported by only one subject were omitted in order to make Figure 74 easier to read. A final reason for these apparent inconsistencies is that participants started at different points; some used the table of contents, while others started by scanning the section titles.

In summary, participants performed the reading task by leafing through the document in a largely linear manner, inspecting every page. They used a set of reading strategies. These act as filters of successively higher cost and fidelity that are applied to a

decreasing pool of potentially relevant material. Thus the participants were not using the 'satisficing rule' and were using strategies that seemed to be attempting to maximise the rate of information gain. Since these two features are key aspects of the strategy used in the best-case model, they support the validity of directly comparing the best case model and expert human performance.

1.2 Task performance

The strategies adopted by participants indicate that performance with paper can be characterised by its low precision and high recall. Participants tended to look through an entire document when searching for information relevant to queries.

Since the almost all participants went through the entire document linearly, the mean traditional precision is low: 0.12. This is somewhat lower than the complementary measure of 'proportion of reading time spent reading relevant' which is 0.36. This shows that participants spent three times as long with relevant paragraphs as with irrelevant ones. Although this might seem self evident it is actually partly a result of the participants' skimming strategies which allow them to adjust the amount of time spent reading a paragraph according to its expected relevance.

In contrast to this low traditional precision and 'proportion of reading time spent reading relevant', the 'proportion of time spent reading' was very high. The mean 'proportion of time spent reading' was 0.94. This reflects both the strategies used and the effectiveness of paper's navigational affordances. The strategies used generally required very little observable navigation, which essentially consisted of a page turn once each page had been read. Hence, participants spent almost all their time actually reading.

The mean interactive precision was 0.7. This is rather higher than the 0.56 obtained in the 'without GridVis' condition of the initial experimental evaluation of GridVis (see chapter 5 section 3.2.2). This confirms the suspicion that participants in the experimental evaluation were moving on as soon as they suspected a paragraph was relevant, rather than actually absorbing the content. Forcing participants to highlight the paragraphs, in this experiment, seems to have made people look at the content more carefully, thereby increasing interactive precision. It also made the task much closer to the kind of goal-directed search done, in everyday work, rather than that done in the experimental evaluation reported in chapter 5 section 3.

The fact that participants were going through the documents in their entirety also meant that the recall was high. The mean overall recall was 0.63, the mean core recall was 0.73 and the mean peripheral recall was 0.47. Although these recall scores are high, there are not perfect. Given that the participants went through the entire documents one could argue that they should have picked up on everything that was relevant. One factor which might contribute the participants missing relevant paragraphs is the variation in what is perceived as relevant (see chapter 5 section 3.1.4). Another factor might be the strategies used. The use of skimming strategies, meant that although most paragraphs were looked at, they were not read in any detail. Hence relevant paragraphs could be missed if, for example, they did not contain the keywords being looked for. It is interesting to note that the core recall is much higher then the peripheral recall. This could also be a product of the type of skimming strategies used. Paragraphs with judge relevance rating of 1 or 2 will be less directly concerned with the query. Because of this there will be fewer indications of its relevance available to the skim reader. For example, the section title might not look relevant, or the paragraph may contain fewer or none of key words being searched for.

1.3 The effect of document length

It was expected that increasing document length would reduce performance. To test this hypothesis Page's L test was used to see whether there were statistically significant trends in this direction. Page's L test is ideal here since it is a non-parametric test for ordered trends in more than two related samples (Neave & Worthington 1988). It tests the hypothesis that conditions are ordered in a specific sequence (i.e. $R_{short} > R_{medium} > R_{long}$,) without the need to assume a normal distribution. Essentially the scores are ranked across conditions (i.e. length), and the mean rank across subjects is then used to calculate the L statistic which in turn is used, along with the sample size, to look up the significance in a statistical table. This test was used, as opposed to a parametric alternative, because when the data was plotted in histograms it did not show a normal distribution.

As can be seen in Figure 75[28], task time increases steadily with document length; this trend was significant to $p<0.01$ ($L=319$, $N=24$).

[28] See Appendix 6 for versions, of the graphs in this section, in which the individual data points are displayed as well as means.

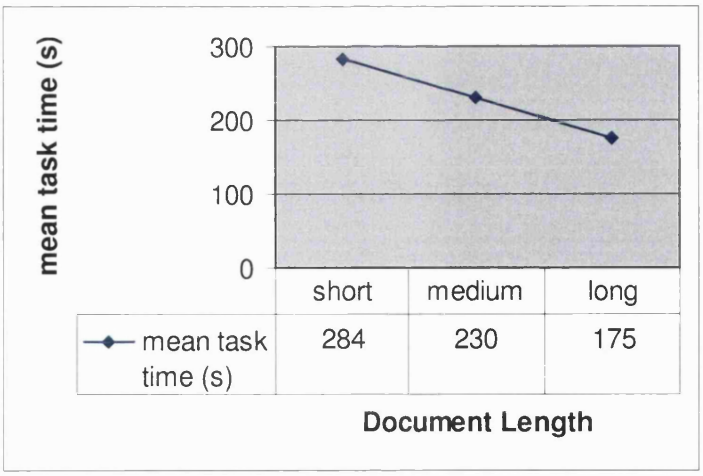
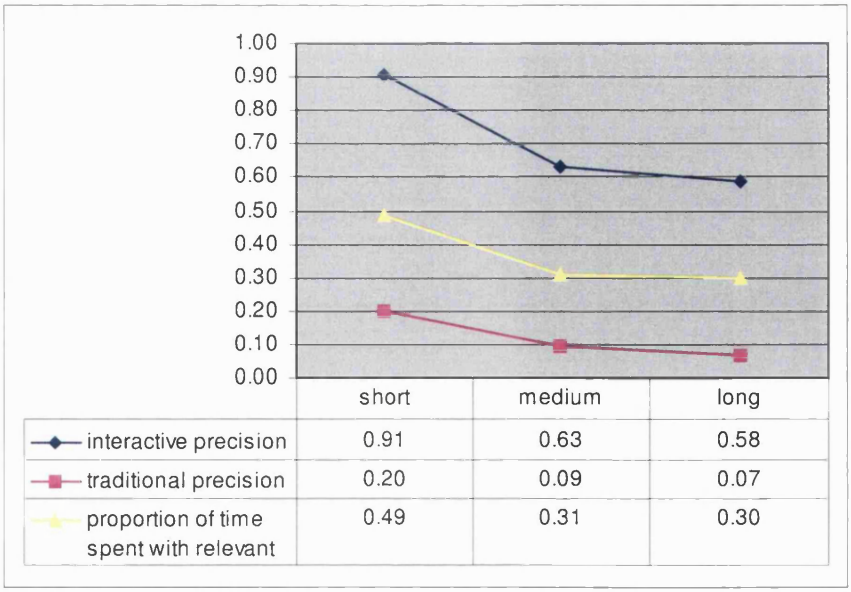


Figure 75 The mean time taken to complete a question for different document lengths.

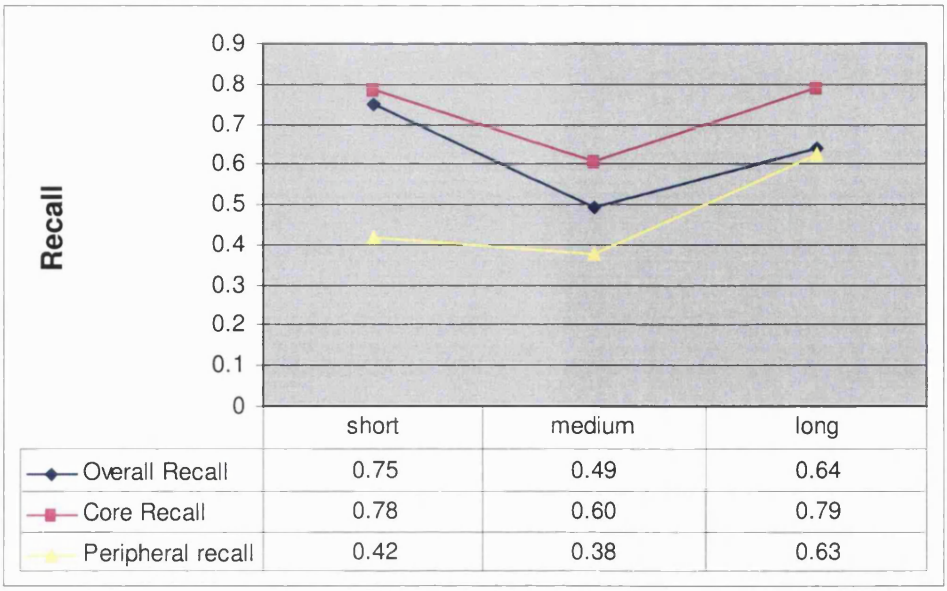
Similarly performance declined with increasing document length on the three precision metrics (see Figure 76); result precision, traditional precision and 'proportion of reading time spent reading relevant'. This trend was statistically significant for all three metrics; for $p<0.01$ for result precision ($L=319, N=24$), for $p<0.01$ traditional precision ($L=214, N=16$), for 'proportion of reading time spent reading relevant' $p<0.01$ ($L=211, N=16$). This effect of length can be explained by the decrease in the proportion of relevant paragraphs which occurred as length increased.



	short	medium	long
interactive precision	0.91	0.63	0.58
traditional precision	0.20	0.09	0.07
proportion of time spent with relevant	0.49	0.31	0.30

Figure 76 The mean result and traditional precision, and 'proportion of reading time spent reading relevant' across the different document lengths.

For recall, the picture is somewhat more complex. Overall, core and peripheral recall have their lowest values for the medium length document and not the long one. Figure 76 shows that for overall and core recall there is no trend but suggests the possibility of a trend for peripheral recall. The trend for peripheral recall is, however, not significant, $p>0.05$ ($L=296$, $N=24$). This result is not that surprising given the strategies used. Participants generally went through the documents in a linear fashion regardless of the document's length. Hence, in terms of recall, longer documents were at no disadvantage.



	short	medium	long
Overall Recall	0.75	0.49	0.64
Core Recall	0.78	0.60	0.79
Peripheral recall	0.42	0.38	0.63

Figure 77 The mean recall across the different document lengths.

Finally for the 'proportion of time spent reading', there was no effect of length. Again this can be explained by the fact that the same reading strategies were used for different document lengths.

Therefore, in summary, document length only had a significant effect on task time, which increased with document length, and on precision, which decreased with document length.

1.4 The effect of question difficulty

Unfortunately, the question difficulty effect did not have a consistent or statistically significant effect on performance; $p>0.2$ for all metrics. The performance metrics were tested using the Wilcoxon signed ranks test. This is a non-parametric test, and thus suitable for this non-normally distributed data, which is designed for detecting differences between two related samples (Howell 1997).

This result was unexpected as the manipulation of question difficulty was based on one used successfully in the evaluation of SuperBook (Egan et al. 1989). The failure of this manipulation to work here could be explained by the different reading strategies used by participants in this experiment and in those in the evaluation by Egan et al (Egan et al. 1989). In the SuperBook evaluation the subjects relied on using the key words given in the query since they were using computerised string searches together with an outline view listing the section headings. Additionally, as the document used was very long (562 pages), the participants relied more heavily on section titles than they did with the comparatively short documents used here (5 to 14 pages).

The adaptation of the procedure for manipulating question difficulty, from the original procedure used with SuperBook, may have reduced its effectiveness. In the SuperBook, evaluation questions were classified as hard when there were no keywords in the heading. Here question difficulty was varied by changing the number of elements, in the headings of relevant paragraphs which related to the question (see section 3.2.1.3 for more detail on this). For easy questions there were two elements, while for hard questions there was only one. Perhaps the manipulation would have been effective if the headings for hard questions had no information relating to the query in them.

2 How does the best-case GridVis performance fare against the benchmark?

This section compares the benchmark performance (i.e. expert human performance on paper), described in the previous section, to that of GridVis as computed by the best-case model.

An important point here is that this research seeks to generalise its conclusions to interactive visualisations of paragraph-level metadata in general. It is possible to argue that any conclusions only apply to GridVis since this is what the best-case model is rooted in. Although GridVis is a particular interactive visualisation of paragraph-level metadata, it is a competent embodiment of the general affordances offered by this class of tool; i.e. the display of paragraph-level metadata and its location, applicability and co-occurrence. As such the conclusions about potential utility of GridVis is gives useful insights for all visualisations of paragraph-level metadata.

The differences between the GridVis and paper performance are tested for statistical significance using the non-parametric sign test. A non-parametric test had to be used because, when plotted in histogram, the benchmark data did not appear normally distributed. The sign test involves comparing the best-case score to the mean score of each subject (Neave & Worthington 1988). The number of subject means larger than the benchmark score are counted. This count is then looked up in a statistical table to obtain significance values. Since this test ignores magnitude information it is less sensitive to small significant differences than a parametric alternative, and thus acts as a more stringent test.

As described in chapter 6, several variations of the best-case model were run, the performance of most of these variations is reported below. Some of these variations are

used to evaluate the trade-offs involved in using the different paragraph selection strategies. To briefly recap, there were four different paragraph cell selection strategies, so the performance of each of these is shown below. A sensitivity analysis examined the best-case model's sensitivity to error in the estimates of cost. Although it has been argued that the default cost estimates are reasonable, human performance times will differ substantially, if not systematically, from them. Hence for the sensitivity analysis all actual and expected costs were adjusted using a single coefficient which resulted in them being, 25%, 50%, 75%, 125%, 200%, and 400% of their default values.

2.1 Recall performance

Recall with the best-case model performance of GridVis is better than that with paper if co-occurrence information is used, and about the same or slightly worse without it. This basic result is illustrated in Figure 78. Overall Recall is significantly higher with GridVis than with paper for both co-occurrence based strategies ($p<0.01, N=24$), and is about the same for non-co-occurrence based strategies ($p>0.05, N=24$). Hence with the use of co-occurrence, overall recall is substantially higher than that achieved with paper.

Comparison of the model's performance with and without the use of a co-occurrence based strategy confirms results of the earlier investigation into the effect of cell selection strategy. This investigation, reported in chapter 5, section 3.3.2.2, predicted an improvement in recall with the use of co-occurrence. However, the improvement found here is much larger; rather than a 10% improvement; it is a 43% improvement. Thus the use of co-occurrence in cell selection can substantially increase recall.

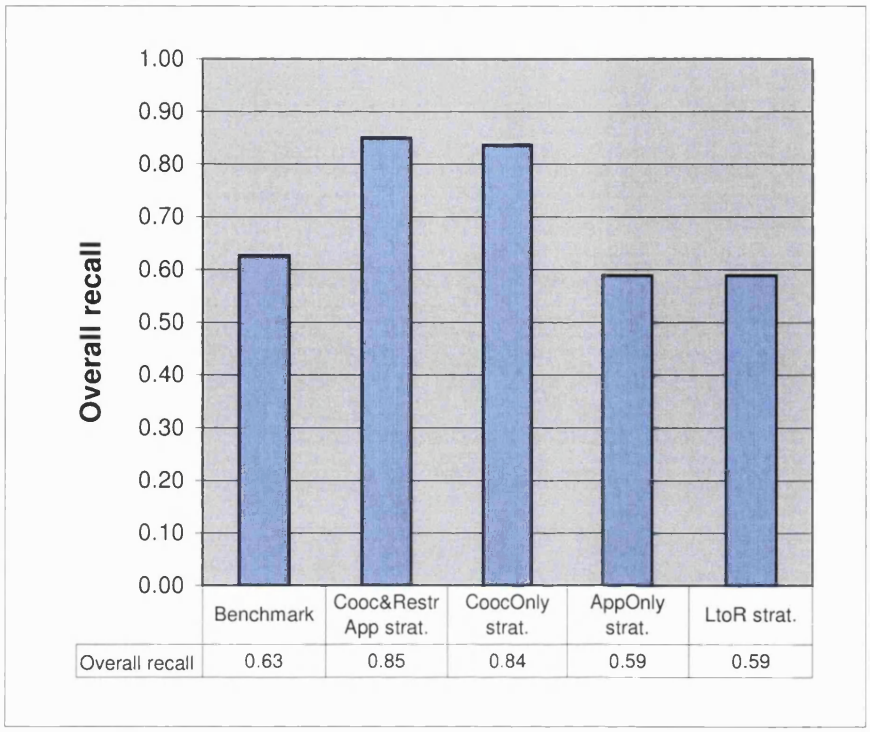
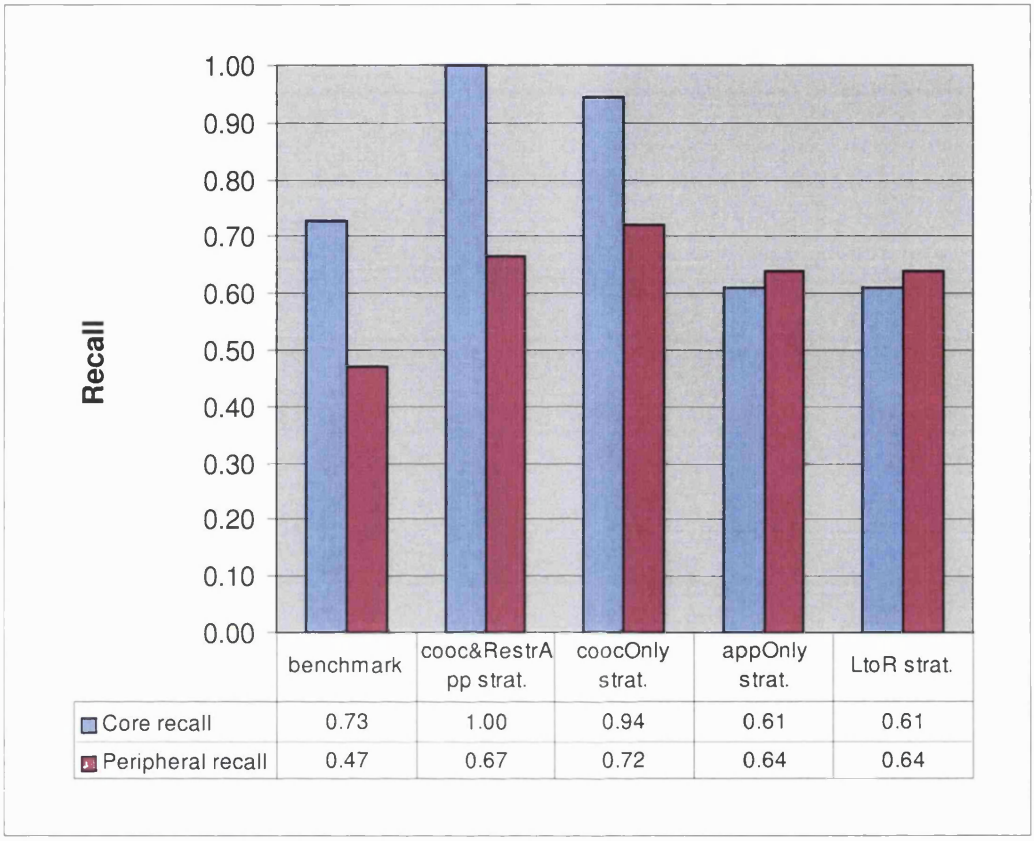


Figure 78 Overall Recall scores

When the core and peripheral recall scores are examined individually (see Figure 79) the advantage offered by the co-occurrence strategies is further elucidated. For paper, the core recall is substantially higher than peripheral recall. On reflection, this result seems characteristic of healthy information foraging; paragraphs of high information value are concentrated upon due to their higher profitability. With GridVis, it is only the co-occurrence based strategies, which exhibit this pattern. With AppOnly and LtoR strategies, there is very little difference between core and peripheral recall. The likely explanation for this is that without co-occurrence information there is not enough information to sufficiently resolve different degrees of relevance; thus time is not consistently invested in high relevance paragraphs. As a result if non-co-occurrence based strategies are used, the core recall is lower for GridVis than for paper ($p<0.05$, $N=24$), while for peripheral recall there is no significant difference between GridVis and paper ($p>0.05$, $N=24$). This is important since core recall is more important than

peripheral recall (see chapter 6, section 4). Thus one could argue that since one cannot successfully concentrate on high value items using the non-co-occurrence based strategies, recall performance is effectively worse using these strategies than it is with paper.



	benchmark	cooc&RestrApp strat.	coocOnly strat.	appOnly strat.	LtoR strat.
Core recall	0.73	1.00	0.94	0.61	0.61
Peripheral recall	0.47	0.67	0.72	0.64	0.64

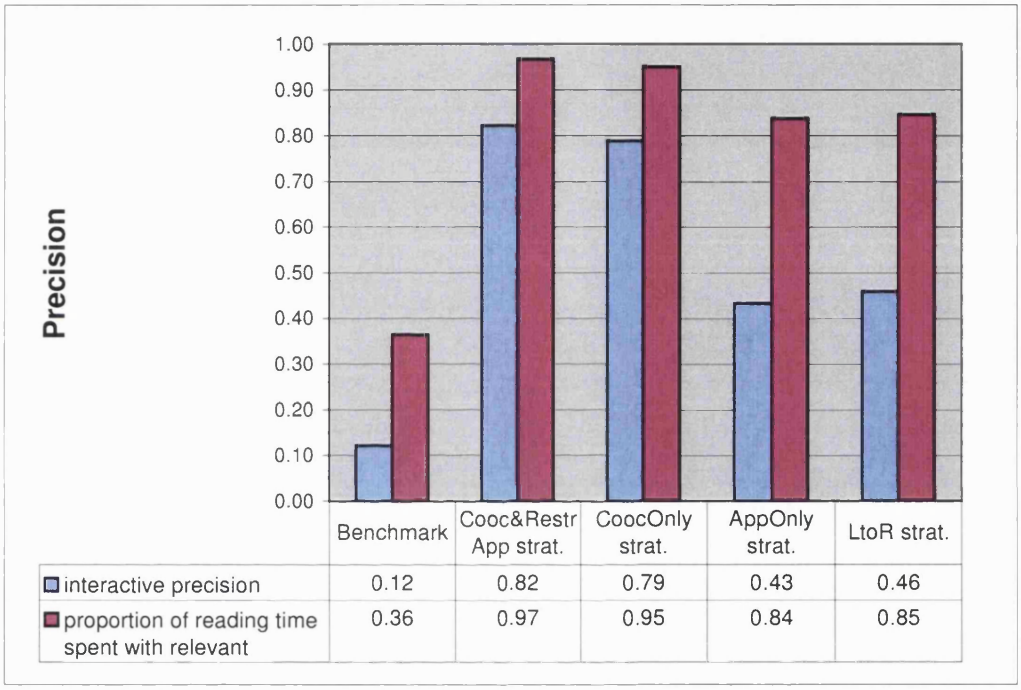
Figure 79 Core and Peripheral recall scores

The use of applicability information does not seem to lead to any improvement in recall scores. Scores for the AppOnly strategy are the same as those for the LtoR strategy. Similarly the scores for the 'co-occurrence and applicability' strategy are little different from those of the 'co-occurrence only strategy'.

The sensitivity analysis showed that predictions of recall were not dependent on time costs. Costs of 25% to 400% of the original estimates had no substantial or systematic effect on recall scores.

2.2 Precision performance

Since GridVis was designed to reduce the amount of irrelevant material read, one would expect precision scores to be higher than those for paper. This is what was found (see Figure 80). The traditional precision was very good; much higher then paper for all conditions ($p<0.01, N=16$). The 'proportion of reading time spent reading relevant' measure reinforces this conclusion ($p<0.01, N=16$).



	Benchmark	Cooc&Restr App strat.	CoocOnly strat.	AppOnly strat.	LtoR strat.
interactive precision	0.12	0.82	0.79	0.43	0.46
proportion of reading time spent with relevant	0.36	0.97	0.95	0.84	0.85

Figure 80 Precision performance scores

As with recall, the use of co-occurrence information substantially improves precision scores, but applicability seems to have no effect. In addition, the sensitivity analysis found that precision was not dependent on time costs. Costs of 25% to 400% of the original estimates had no substantial or systematic effect on the precision scores.

2.3 Task time performance

The results shown in Figure 81 reveal the trade-off between the higher time investment required by co-occurrence based strategies, and the higher recall and precision they afford. The co-occurrence based strategies result in about 40% slower task times than paper, while the non-occurrence based strategies result in about 40% faster times then paper ($p<0.01$, $N=24$). So although the non-co-occurrence based strategies may have recall which is slightly worse than that of paper, they are substantially faster.

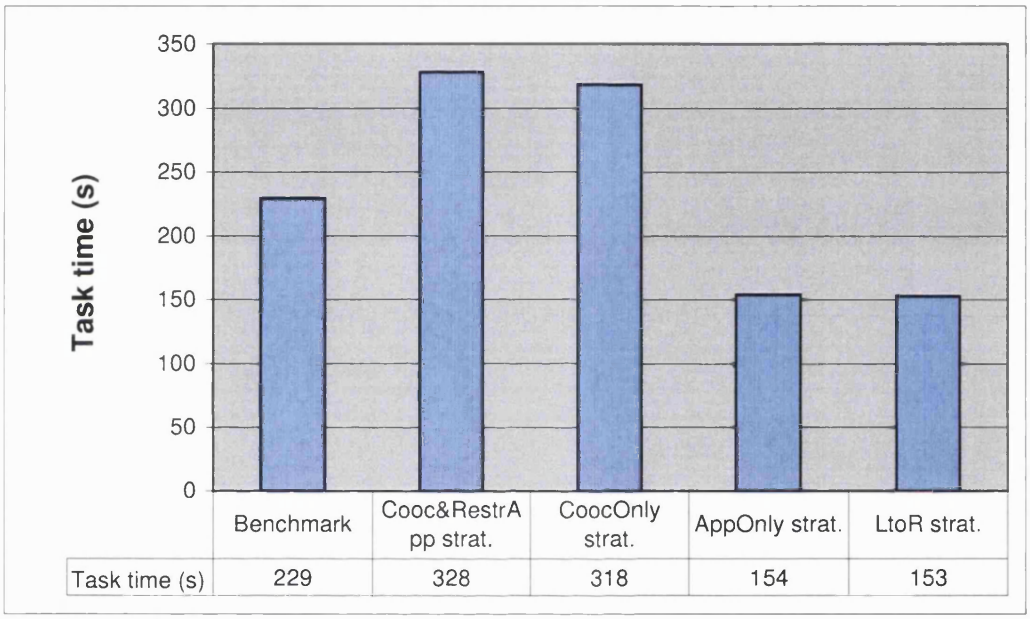
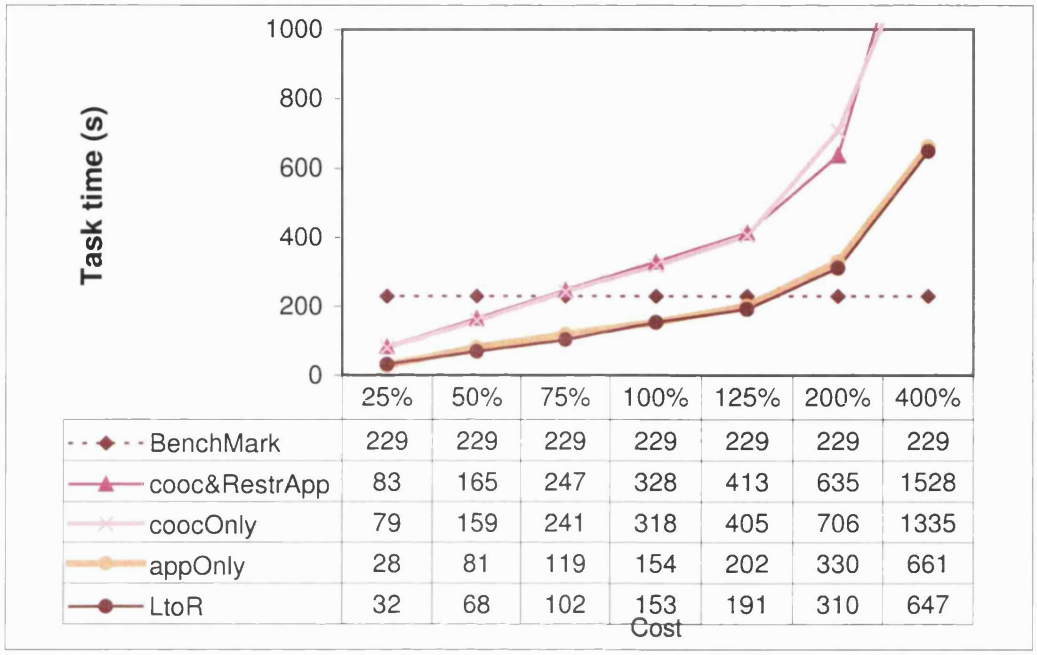


Figure 81 Task time performance

Evidently, the task time is very much dependent on the estimates made of the time it takes to operate GridVis. The results of the sensitivity analysis are shown in Figure 82. They suggest that the trade-off, relative to paper, between task time and retrieval performance only hold for cost estimates that are within 25% of the default estimates in the best-case model. If the costs fall below these, then GridVis is faster than paper, whichever strategy is used. If the costs rise above them, than GridVis is slower than paper, regardless of the strategy used.

The finding that task time sinks below that of paper could be used to set design goals for improvements to GridVis; any changes that bring 50% reduction in usage times would make GridVis substantially faster than paper even if co-occurrence strategies were used.



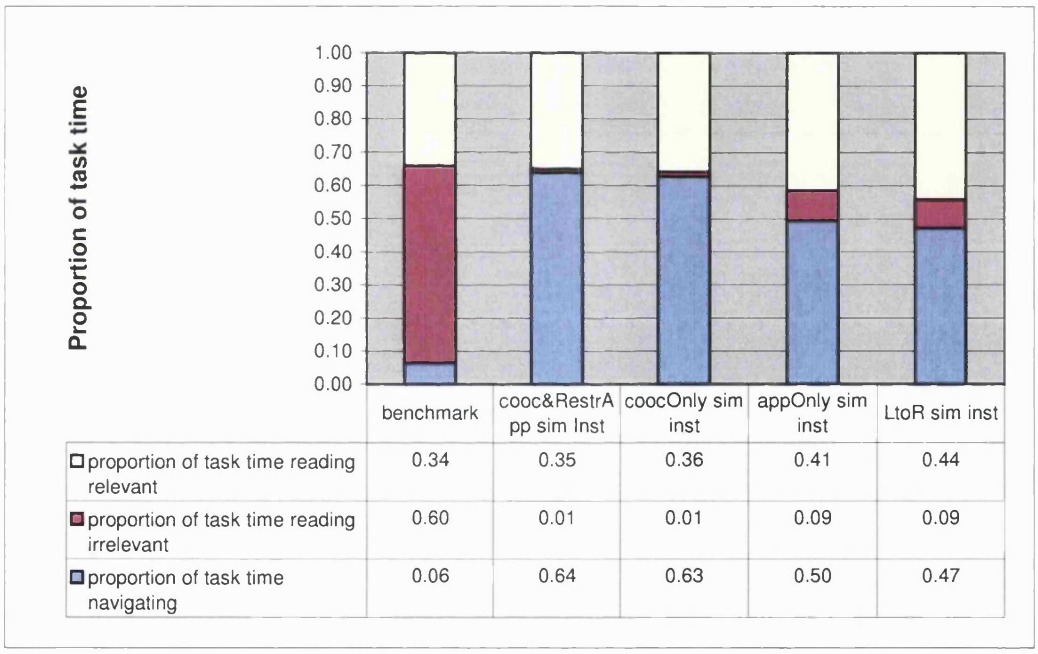
	25%	50%	75%	100%	125%	200%	400%
BenchMark	229	229	229	229	229	229	229
cooc&RestrApp	83	165	247	328	413	635	1528
coocOnly	79	159	241	318	405	706	1335
appOnly	28	81	119	154	202	330	661
LtoR	32	68	102	153	191	310	647

Figure 82 The dependence of task time on the cost estimates for GridVis usage. The results for paper are not part of the sensitivity analysis, but are included to enable comparison.

2.4 Does the extra cost of navigation with GridVis outweigh the cost savings of the paragraphs avoided?

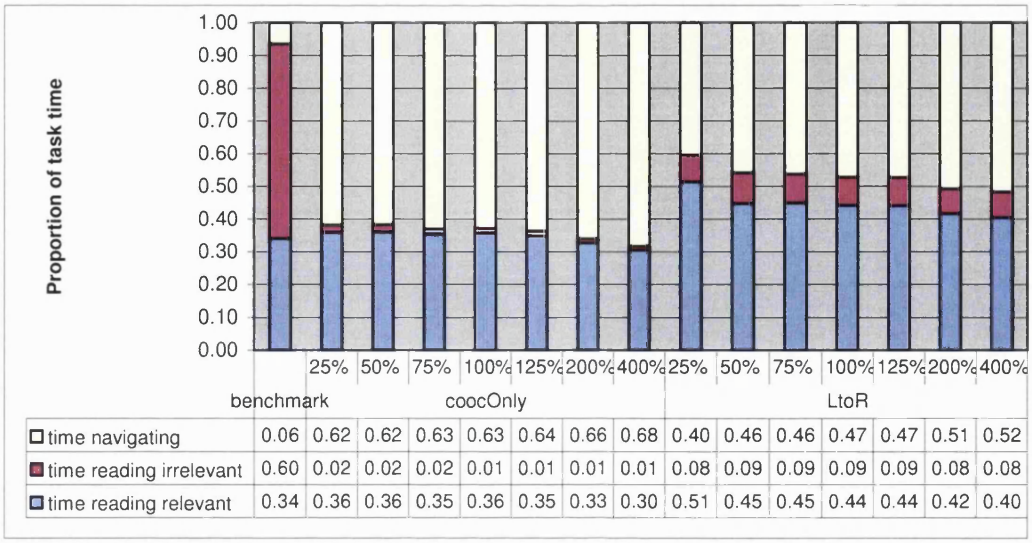
One of the assumptions that underlie the idea behind GridVis is that it will save the user time by reducing the need to read irrelevant paragraphs. As discussed in the previous section, the use of paragraph level metadata results in a more precise search. Hence fewer irrelevant paragraphs need to be read and time is saved. However, if GridVis imposes too high a navigation cost, then this saving will be lost. Figure 83 shows that GridVis does indeed impose a large navigation cost relative to paper, and that this

increases with the amount of information used in the paragraph selection strategy. It also shows that GridVis much reduces the time spent reading irrelevant material, reducing it to 1-2% if the co-occurrence based strategies are used. However to a large degree the advantage of avoiding irrelevant material is largely cancelled out by the extra time invested in navigation. Thus although the non-co-occurrence based strategies allow a significantly higher proportion of time to be spent reading relevant material ($p<0.05$, $N=16$), it is only an increase of 8 percentage points. The sensitivity analysis showed that changes in cost estimates only have a minimal impact, leaving the basic conclusion intact (see Figure 84). Hence, in sum, the navigation costs imposed by GridVis largely cancel out any benefits accrued by avoiding irrelevant material.



	benchmark	cooc&RestrApp sim Inst	coocOnly sim inst	appOnly sim inst	LtoR sim inst
proportion of task time reading relevant	0.34	0.35	0.36	0.41	0.44
proportion of task time reading irrelevant	0.60	0.01	0.01	0.09	0.09
proportion of task time navigating	0.06	0.64	0.63	0.50	0.47

Figure 83 Scores for the proportion of time spent reading and navigating



	benchmark	coocOnly							LtoR						
		25%	50%	75%	100%	125%	200%	400%	25%	50%	75%	100%	125%	200%	400%
time navigating	0.06	0.62	0.62	0.63	0.63	0.64	0.66	0.68	0.40	0.46	0.46	0.47	0.47	0.51	0.52
time reading irrelevant	0.60	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.08	0.09	0.09	0.09	0.09	0.08	0.08
time reading relevant	0.34	0.36	0.36	0.35	0.36	0.35	0.33	0.30	0.51	0.45	0.45	0.44	0.44	0.42	0.40

Figure 84 Results from the sensitivity analysis of the proportion of times spent reading and navigating

2.5 Does GridVis increase the quality and relevance of material found?

It was hoped that GridVis would offer a better, or at least not significantly worse, ability to locate relevant material. This was the second of the two assumptions that underlie the idea behind GridVis and paragraph level metadata in general. The results reported at the start of section 2 suggest that this assumption is largely correct. For co-occurrence based strategies both recall and precision are substantially higher than that found with paper. For the non-co-occurrence based strategies the precision is better than paper's while overall recall and peripheral recall is not significantly different. It is only for core recall with non-co-occurrence based strategies that GridVis's performance is significantly worse than paper's (0.61 vs. 0.74). Hence GridVis offers recall and precision that are generally better than those obtained with paper, with the exception of core recall when non-co-occurrence based strategies are used.

2.6 The trade-off – balancing task time and retrieval effectiveness

Essentially, these results reveal a trade-off between task time and retrieval effectiveness. Either GridVis offers better retrieval performance than paper, at the cost of longer task times, or offers faster performance but with slightly worse recall. This suggests particular conditions in which a user might want to use GridVis instead of paper.

However what if a user was concerned about both time and recall? Which tool would they choose? According to information foraging theory they would choose the most profitable tool, i.e. the one which produced the most valuable information per unit time. When the profitability, or rate of gain, for paper and GridVis (see Figure 85) are compared, is clear that there is little difference ($p>0.5$, $N=16$). This suggests that if time and retrieval considerations are taken together paper and GridVis are roughly equivalent. Thus it is the trade-offs between task time and retrieval effectives that are important in determining when GridVis might be advantageous.

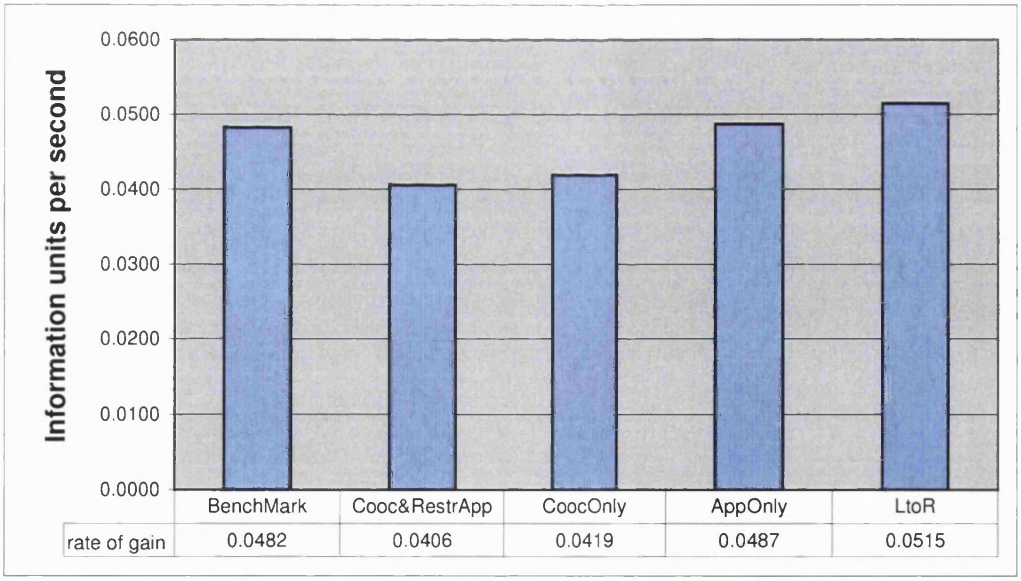


Figure 85 The rate of gain or profitability (information units per second).

However in the sensitivity analysis it is clear that as the cost of using GridVis decreases, the rate of gain steadily rises above that of paper (see Figure 86). Thus perhaps for particularly adept users, or for all users if the design is substantially improved, GridVis will offer superior overall profitability to paper.

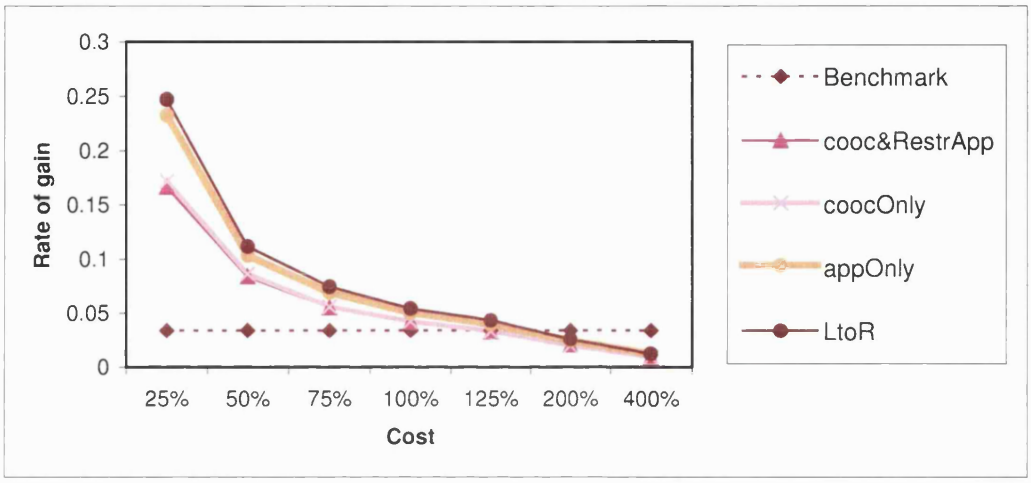


Figure 86 Results from the sensitivity analysis for the rate of information gain

2.7 The effect of document length

Document length seemed to have very little systematic effect on the best-case GridVis performance. The only measure which showed a trend was task time which increased with document length for the co-occurrence based strategies; however according to Pages L test this trend was not significant ($P>0.05$, $L=51$ & $N=4$ for co-occurrence strategies alone, $L=93$ & $N=8$ for all strategies).

This is an encouraging result, since it suggests that the performance of GridVis is independent of document length. In section 1.3 it was shown that the opposite is true for paper documents. With paper, as the document's length increases, task time increases while precision decreases.

Nevertheless, GridVis' independence from document length may not hold outside of the range examined, i.e. 5 to 14 pages. However, as found in chapter 3, section 2.2.1, such documents are relatively rare in the application area studied.

2.8 The importance of exploitation cost prediction

The process of modelling GridVis has demonstrated that the fidelity of estimations of exploitation costs and information scent is crucially important in information foraging models. In an early incarnation of the best-case model expected exploitations costs were estimated solely on the basis of scent. This resulted in cost estimates with a very low correlation to actual exploitation costs. As a result the model's behaviour was very different from that leading to the results reported above, and far from optimal. Improvements were made to exploitation cost estimation, by incorporating additional cues, which much improved its correlation with actual exploitation times. The final model of exploitation cost estimation produced correlations to actual exploitation costs between 0.95 and 0.26, with the mean being 0.65. Given the potential distortions that erroneous models of expected cost can introduce to foraging, it is notable that the degree of fit between estimated and actual costs has not been reported in previous information foraging work (Pirolli & Card 1999). The difficulties associated with producing good models of expected exploitation cost are discussed in section 3.2.

This demonstrates that the accuracy of information foraging models is dependent on the fidelity of the models of expected exploitation costs and expected scent. The size and distribution of deviation between expected and actual values for cost and scent will determine the optimality of the foraging decisions and thus the level of performance. Furthermore it suggests that cues which indicate exploitation costs are just as important as those bearing information scent because if the forager cannot assess exploitation cost they cannot make rational foraging decisions. If further work confirms the importance

of cost cues, it could form the basis of a design guideline for information visualisations. This is discussed further in chapter 8, section 2.3.

3 General discussion

The results presented in this chapter suggest that GridVis, and thus interactive visualisation of paragraph-level metadata in general, does have the potential to support goal-directed search within documents. However it only has the potential to exceed the support offered by paper documents in certain ways. An expert user of GridVis can do goal-directed search faster, with recall which is approximately the same or slightly worse than that of an expert user of paper. Alternatively they can use GridVis to get higher recall, but take more time. Thus GridVis can potentially be used to either give an advantage in terms of time or in terms of recall, but not both at once.

An important question is to what extent these benefits will translate to real world human performance. Given the nature of the best-case model it is likely that performance in the real work place use of GridVis would be somewhat lower than that of the model. The difficulty of determining the nature of this reduction is one of the disadvantages of best-case analyses. This is discussed further in chapter 6, section 1.1.2. Nevertheless, it is instructive to consider the magnitude of the potential advantages uncovered. The best-case model suggests that task time can be reduced by 40% or recall can be increased by 40%. Even a halving of these gains would leave GridVis with a substantial advantage over paper. Also, given that it is only with paper that document length has a deleterious effect on performance, this advantage would be greater for longer documents. This suggests that GridVis, or more generally, the interactive visualisation of paragraph-level metadata, will benefit users who place a premium on either task time or recall.

These results also reveal a crucial drawback of GridVis. They suggest that GridVis only has an advantage over paper for users who value time over recall or vice versa. When time and information are considered together paper performance is about the same as GridVis best-case performance. Given that real-world performance with GridVis will be

lower than its best-case performance it is likely that, in this respect, paper will fare substantially better. Thus a user might want to use GridVis for tasks where time or recall alone are particularly critical, and paper otherwise (see Figure 87).

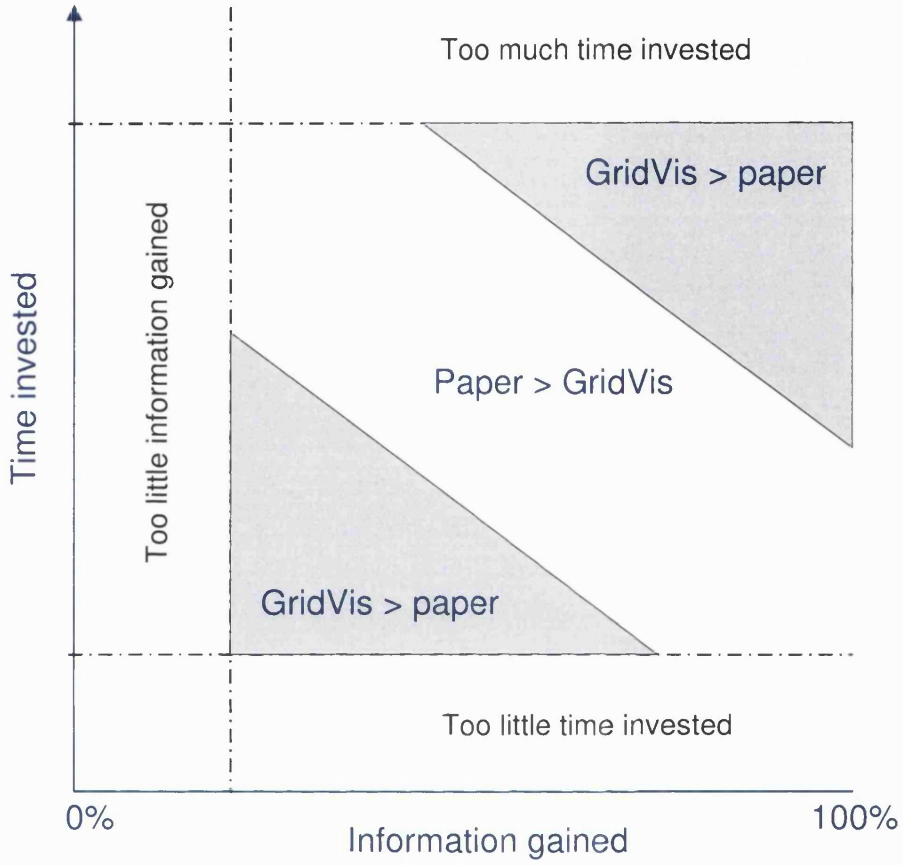


Figure 87 An illustration of when paper or GridVis is advantageous.

The fact that GridVis is only advantageous in particular circumstances is itself a major drawback. It means that GridVis, or more generally paragraph level metadata, is only useful for a proportion of goal-directed searches. One could argue that factors other than performance such as user satisfaction might mitigate against this failing. However, without the promise of increased productivity, it would be difficult to justify the cost of high cost metadata production as well as that of deployment and training needed if this sort of tool were to be deployed in a corporate setting.

3.1 Design implications

The results have a number of implications for the design of interactive paragraph level metadata for goal-directed search. Firstly they show that co-occurrence information can be extremely effective in increasing precision and recall. Thus the emphasis in GridVis on facilitating its use has been vindicated. Secondly they show that applicability information is of little use, and thus might be safely excluded from the visualisation to reduce the complexity of the display and simplify metadata production. A third, more tentative point, is that goal-directed search can be given superior support throughout the space in Figure 87, *if* the visualisation design can be improved such that it is substantially, say 50%, faster to use (see section 2.6). Finally, the process of constructing the model revealed the importance of cues which indicate expected exploitation costs (as discussed further in section 3.2 below). Thus the present dynamic level-of-detail management feature (see chapter 4, section 3.2.3.5) (which was not modelled) is problematic, since it makes it impossible to determine the number of paragraphs on a tag.

3.2 Critical assessment

This section takes a critical look at the best-case analysis and the modelling techniques it employed. The best-case analysis method developed and applied in this thesis allowed the potential performance of a new tool to be compared against that of the competition. Its major drawback is that it cannot specify the degree to which real world performance

will fall short of the best-case performance[29]. This weakens the conclusions that can be made when performance is superior to than that of a benchmark. However it does allow numerous parameters, such as paragraph selection strategy and differing usage costs, to be explored. The economy in the number of participants it afforded allowed a robust evaluation of GridVis in a corporate setting where the resources available to academic research were limited. A more detailed discussion of the general advantages and limitations of best-case analyses can be found in chapter 6, section 1.1.2. Here, the modelling techniques used to create the best-case model, namely GOMS and information foraging theory (IFT) are subjected to scrutiny.

GOMS, or more precisely, NGOMSL, was used to predict the time taken to use GridVis optimally. As such, the best-case performance of GridVis reported here rests on the validity of the GOMS models. The variant of GOMS used here, NGOMSL, has been previously found to give accurate predictions of performance time on a number of different tasks (Kieras 1996). However finer grained techniques such as CPM-GOMS might have produced more accurate predictions (John & Kieras 1996).

The absence of a treatment of satisficing is a limitation of information foraging theory in general and the optimal diet selection algorithm in particular. In a review of the literature on menu selection Norman notes that the satisficing rule, i.e. termination of the search when a 'good enough' item being found, is a common strategy (Norman 1991). Miller and Remington use a satisficing rule in their model of web site navigation arguing that it requires fewer cognitive resources than an optimising strategy, where all

---

[29] On first reflection a comparison of results from the best-case with the much lower results observed in the empirical evaluation (Chapter 5, section 3.2.2) would suggest that this is a serious problem. However the poor performance in the experiment is largely due to the difficulty of the documents used, as evidenced by the equally poor performance seen without GridVis. Also there were various design problems with the version of GridVis used for experiment, the participants were not expert users of the device, and were not familiar with the taxonomy. Thus they visited the same paragraphs repeatedly, made little use of co-occurrence and could not have used the 'expert' strategy for tag selection.

alternatives are compared (Miller & Remington 2001). The use of information foraging theory in this research provided a solid theoretical basis for modelling information seeking behaviour, however it meant that a satisficing rule was not modelled. However, the absence of a satisficing rule in the best-case model of GridVis use seems appropriate since it was also absent in the readers use of paper; participants in the paper performance experiment did not stop once they had enough information - they proceeded linearly through the entire document (see chapter 7, section 1.1).

There is a limitation in the approach used to estimate scent and expected cost. This is noteworthy since foraging decisions are largely dependent on these estimates. The rational analysis approach used ignores what might be important systematic error in human estimates of cost and scent. For example, exploitation costs were modelled by fitting a function incorporating information from the visualisation to actual costs. This produced estimates for exploitation time with some systematic errors. For example the tag exploitation time for very high scent items are underestimated for the co-occurrence based strategies. This is because the estimate is based on a figure for the mean number of selections, and a larger than average number of selections are made on high scent items. These types of systematic error stem from the particularities of the function used to make the estimate. Thus it could be argued that the foraging decisions which result from these estimates are distortions. However it is unlikely that human estimates will track actual values with a purely random error. The estimates are bound by the information available and the manner in which human cognition can use it; thus some systematic error is highly likely. It is hoped that the model of exploitation cost reproduced these errors since it was constructed to incorporate as much of the available information as possible while respecting the limits of human cognition. Nevertheless the present research does not offer any guarantees that any systematic errors in human

estimates are reproduced in the foraging model. Previous information foraging models have paid scant attention to how cues in the interface are used to estimate expected exploitation cost and what errors this may produce (Pirolli & Card 1999). Further research on this matter is therefore needed (see chapter 8, section 2.3).

There is a further, and rather subtle, problem with the estimation of information scent. In the best-case, information scent tracks relevance as defined by the relevance score aggregated from the scores of three expert judges (see Appendix 2). Therefore for the model there is a one-to-one correspondence between the scent, and the information value upon which performance is scored. In the paper experiment this is not quite the case. As has been seen previously in this research (chapter 5, section 3.1.4), relevance judgements vary substantially from person to person. As a result scent for individual participants will have a less than perfect correspondence to information value as defined by the judges. Thus the scoring of performance will somewhat favour the best-case model since, in a sense, it has a better understanding of what is required.

A number of other limitations of the best-case model are discussed elsewhere. The simplistic nature of the measures cost and information used are discussed in section 2.2.1. The limitations involved in the use of perfect information scent are discussed in chapter 6, section 2.2.2.1. Possibly relevant features and phenomena not included in the model are discussed in chapter 6, section 2.1.3.

4 Chapter summary

Analysis of data from the paper performance experiment allowed reading strategies to be described. This suggests that the use of paper is comparable to the use of GridVis embodied in the best-case model; neither use strategies that follow the 'satisficing rule' and both attempt to maximise the rate of information gain. The best-case model of

GridVis produced results which show that relative to paper GridVis can offer either higher recall or faster performance. However if value is placed both on the speed and recall, then GridVis can only equal paper's performance. Increasing document length was found to increase task time and decrease precision for user of paper, but to have no effect with GridVis. The impact of these results on the possibility of the use of interactive visualisations of paragraph-level metadata in the work place is discussed and the implications for future improvements to the design are outlined.

Chapter 8 Conclusions and further work

1 Conclusions and contributions

This section summarises and ties together the work done in this thesis. A short section is devoted to each major contribution. Each of these is comprised of a description of the major contribution as well as the related conclusions and minor contributions. Table 26 lists the contributions made, alongside icons which are used to indicate where they are described in section 1.

An approach to visualising within-document metadata which facilitates goal-directed search within-documents	1
A characterisation of the documents suitable for enrichment with paragraph-level metadata	1
A new method for soliciting metadata and metadata taxonomies that are focused on the information needed to discriminate relevant material	2
A new evaluation method which allows the ultimate potential of a class of designs to be assessed	2
The application of a best-case analysis to an information seeking tool	3
A characterisation of the high-level strategies used in goal-directed reading	4
The uncovering of theoretical grounds, in information foraging theory, that suggest the importance, for information seeking tools, of cues which communication information on access costs.	5
An evaluation of the utility of the interactive visualisation of paragraph level metadata in the support of goal-directed search within-documents	3

Table 26 The major contributions 1→3 1→5 *and their related minor contributions*

1.1 A new approach for supporting goal-directed search within corporate documents

1 This research introduces, develops and evaluates a new approach to the support of goal directed search within corporate documents: the interactive visualisation of paragraph-

level metadata. This uses a number of visualisation techniques to present semantic metadata, which describes the content of each paragraph in entire document.

This contribution rests upon a number of the pieces of research reported in this thesis. Two studies of document usage in a large UK corporate head office suggested that paragraph-level metadata could be useful for a portion of business documents. The thesis then described the method used to prepare high quality metadata for business documents. This included a novel technique, developed for this thesis, to help mould the metadata and metadata taxonomy to typical user information needs. A new visualisation technique was developed and implemented to explore and evaluate the affordances of paragraph-level metadata. Empirical and theoretical evaluations allowed the effectiveness of these affordances to be assessed and recommendations for future designs to be made.

Chapter three reported two studies of document use at the head office of Sainsbury's, a large UK retailer. These characterised the type of document suitable for enhancement with an interactive visualisation of paragraph-level metadata. The studies uncovered evidence for the feasibility of such a tool for supporting goal-directed search. Characterisation of reading strategies and document reuse indicates that much document use is goal-directed. The studies found that over half of the goals readers had for documents could be directly met by metadata. They also revealed that information relevant to reader goals was often in multiple non-contiguous locations, so that it could be usefully brought together by paragraph-level metadata. These two studies thus provided evidence for the feasibility of the support for goal-directed search within corporate documents using interactive visualisation of paragraph-level metadata.

In order to test the utility of paragraph-level metadata it was necessary to ensure it was of high quality. The metadata thus needed to be produced manually. The methodology

used to do this is described in section chapter 4, sections 2.2 and 2.3. It is based on methods developed for producing formal ontologies. These methods were substantially extended using a new technique developed in this research. In this technique potential users were asked how they distinguish the material they would read from that they wouldn't. This allowed the metadata and taxonomy to be constructed precisely around those elements of information required to differentiate relevant material.

The design of an interactive visualisation of paragraph-level metadata is described in chapter 4, section 3. It uses techniques drawn from the information visualisation literature. This literature is reviewed, in chapter 2 section 3, drawing on empirical evaluation where possible to pick out the best performing techniques for integration into the design.

Empirical and theoretical evaluation is brought to bear on, GridVis, the resulting tool, allowing the various affordances it offers to be assessed. The best-case evaluation shows that the use of co-occurrence information can deliver a substantial increase in both recall and precision but that its use almost doubles task time (chapter 7, section 3). This suggests that the utility of GridVis could be much improved by design improvements which reduce the cost of using co-occurrence. It also shows that applicability information has no positive impact on task performance and can thus be dropped from future designs. User trials and empirical evaluation identified the importance of history visualisation in reducing unintended revisits to paragraphs. The process of building the best-case model, and in particular the expert user strategy (chapter 6, section 2.1) allowed design flaws to be identified. This suggested that the standard vertical indented folding tree visualisation should more clearly mark out the sequential grouping resulting from the taxonomy. It also suggested that the level-of-detail management feature was problematic because it obscured how many paragraphs

were on a tag. Thus, this research has allowed the utility of visualisations of paragraph-level metadata to be assessed and has shown what needs to achieved if future designs are to promise better performance than paper.

1.2 A new evaluation technique

In the face of various challenges, described in chapter 5, section 3, a new evaluation technique was developed. This is called 'best-case analysis'. It reveals the potential performance of a class of designs relative to that of the most prevalent or effective competing tool.

From a traditional empirical evaluation of GridVis it became clear that existing approaches to evaluation were not able to determine the functional viability of the basic design behind a tool. This was particularly the case since the tool was designed for use, and in situ evaluation, in a work setting where the resources available to academic research were limited (chapter 5, section 4). For these reasons a new evaluation technique, best-case analysis, was developed.

2 A best-case analysis, described in chapter 6, essentially compares the best-case performance of the tool in question with that of a benchmark (the most prevalent or effective competing tool). The best-case performance of the tool being evaluated is computed using a theoretical model of the tool's use. The benchmark performance is either measured empirically using expert users, or computed using modelling techniques. This new technique is applied to assess the potential of the interactive visualisation of paragraph-level metadata.

3 A detailed dynamic model of the best-case GridVis use is constructed (see chapter 6, section 2) using information foraging theory (Pirolli & Card 1999). This demonstrates the viability of the best-case approach. In addition, it demonstrates a number of

modelling techniques which can be applied to construct best-case models for information seeking tools in general.

The process of modelling uncovered the importance of the hitherto ignored correspondence between estimated and actual exploitation costs (see chapter 6, section 2.2.3.2). This suggests that information about access costs can be important in allowing users to optimise the use of GridVis and information access tools in general.

Results from the paper performance experiment, an experiment on readers performance using paper documents reported in chapter 7, section 1.1, which was chosen as the benchmark, supported the validity of comparing the performance of the benchmark and the best-case model of GridVis. Paper was chosen as the benchmark tool because the literature shows that its many navigational affordances make it the de facto means of finding information in corporate documents (chapter 2, section 4.3.2). The benchmark performance was measured in an experiment where experts were asked to search for information in paper documents (chapter 6, section 3). This showed goal-directed reading to be a configuration of progressively higher fidelity and higher cost strategies, which seem to be applied in accordance with information foraging theory (chapter 7, section 1.1). The similarity of this feature of performance, to that incorporated into the best-case model reinforced the validity of a comparison between the performance of the benchmark and best-case model of GridVis.

1.3 Evaluation of the functional viability of interactive visualisations of paragraph-level metadata

The best-case analysis allowed the functional viability of interactive visualisations of paragraph-level metadata to be evaluated. It showed that GridVis can substantially outperform paper in terms of either task time or search performance (measured by recall

and precision). Thus, GridVis might be of interest to users who put a premium on task time *or* on retrieval performance. However, when both task time and retrieval performance are valued, GridVis can only do *as well as* paper. Thus, GridVis can only offer superior performance for a portion of use scenarios. As a result rolling it out for use in a corporation is unlikely to be worth the cost of deployment, training and metadata production.

A sensitivity analysis examined performance in the hypothetical case where the costs of using GridVis are substantially reduced. This suggested that if the costs of using GridVis were reduced by 50%, or more, GridVis could offer substantially better performance than paper across the board. This is likely to require substantial changes to the current design. As they stand interactive visualisations of paragraph-level metadata do not have the potential needed for widespread use in a corporate environment.

2 Further work

This research has developed and evaluated a new type of tool for within-document search, it however leaves and uncovers an number of avenues for future research.

2.1 Evaluation – alternative scenarios of usage

The inevitable result of bringing a manageable focus to this research has been to select an important subset of information seeking scenarios in which to test GridVis. Further research examining the use of interactive visualisations of paragraph-level metadata could explore a number of alternative scenarios.

Interactive visualisations of paragraph level metadata may also be useful in other types of information seeking. It remains to be seen whether it can aid in the process of query formulation. As argued in chapter 2, section 4.1 the presence of a metadata taxonomy tailored to the user populations' information needs may help the transition from the unexpressed visceral need to compromised need. There are also other, less prevalent, but potentially important, types of search such as information encountering (Erdelez 1996), which could be supported by the interactive visualisation of paragraph-level metadata.

Unlike previous evaluations of with-document search tools (see chapter 2, section 5.1.1) a value was placed on the workplace commodity of time; subjects were asked to work rapidly, the best-case model placed a value on time. However, another approach is to look at performance with hard deadlines since it is quite possible that GridVis has a particular advantage over paper for short deadlines. A best-case model of GridVis use with deadlines, which incorporates the notion of scent, could be built using dynamic programming (Pirolli 1998).

GridVis and the various strategies for using it were tested in isolation; however it would be possible to use them in combination. As discussed in chapter 6, section 2.1.3, a user could switch between paragraph selection strategies, as well as between GridVis and alternative navigation tools such as the scroll bar. This switching could be modelled with IFT and a model which estimates the profitability of each alternative for a given action.

The news digest documents used in the best-case analysis were largely devoid of narrative structure. As a result there was no metadata included to describe structure. However the research reports used in the experimental evaluation were rich in narrative structure, and the frontier users felt that the metadata for these documents which captured narrative information (e.g. 'conclusion', 'evidence') was extremely useful. The utility of this narrative metadata was perhaps not evident in the empirical evaluation due to the general difficulty of the documents. Future work on visualisations of paragraph-level metadata could therefore look at the effectiveness of structural or narrative metadata.

2.2 Paragraph-level metadata and its visualisation

The metadata and metadata taxonomies developed here were very specific to the document types being used. Future work on the production of paragraph level metadata and the taxonomies that house it, should examine the degree to which taxonomies can be generalised or co-opted into use with new document types. This work could build on existing proposals for the management of generalisation, refining and integration of multiple taxonomies (Dourish, Lamping & Rodden 1999).

The evaluation results reported (chapter 7, section 3.1) above indicate a number of approaches to improving the design of interactive visualisations of paragraph-level

metadata. They demonstrated the effectiveness but relative slowness of using co-occurrence information with the current design. Future visualisations might profitably seek to lower the cost of co-occurrence use, or perhaps investigate the utility of the kind of co-occurrence automation feature explored in chapter 4, section 3.2.3.5. One potential approach to lowering the cost of using co-occurrence use would be to present co-occurring tags using rapid serial visual presentation (RSVP) (de Bruijn & Spence 2000). This should be faster then requiring the user pick out the highlighted tags from the display and might remove the need for the grid, allowing for a more compact visualisation. Experience with models of expected exploitation cost suggests that cues which indicate access costs are important. Perhaps the design could be improved by making such cues more accessible by, for example, integration, into the metadata tree, of information on the number paragraphs on a tag and their relative length. In the user study from which the expert tag selection strategy was developed it became clear that users made use of the sequential grouping rather than the hierarchical structure. Research could usefully be done on when horizontal folding trees are used in this way, and on what representations will best support this mode of use. In addition, a problem with the current representation is that for longer documents the taxonomy is too long to fit on a screen. An improved design could overcome this problem with the use of a fisheye technique such as that used for pull down menus (Bederson 2000) as discussed in chapter 2 section 3.2.3.

2.3 *Information foraging models*

The process of building an information foraging model revealed a number of issues which require further research.

The best-case model calculated scent for metadata by summing the information value available through the piece of metadata in question. It was argued that this approach

was more appropriate in the context of this research than the previously used calculation of scent based on word frequency statistics (chapter 6, section 2.2.2.1). Briefly, this was because the small number of words included in the metadata would not allow techniques based on word frequency statistics to produce the high performance required to calculate the best-case performance. The use of this 'perfect scent' begs the question of how human perceived scent differs from this. Recent attempts to model web browsing found that a good fit to aggregated human data could be obtained by adding a 40% Gaussian error to perfect scent (Miller & Remington 2001). However, no justification is given for the assumption that the error in perceived scent is normally distributed. Research is needed to examine the nature of error in scent. By illuminating the nature and cause of these errors such investigations could have useful implications for design.

Similarly, the issue of the nature of human error in the estimation of exploitation cost needs further research. In the process of constructing the best-case model it became clear that the degree of error in the estimation of exploitation time had a large influence on search performance. Research is needed to examine the role of information on access costs in human foraging decisions. If rational use of this information leads to distorted estimates of cost, are these distortions evident in the resulting human foraging behaviour? Furthermore, if cost information does influence foraging decisions, how does it do so? In addition, in what way is the use of cost information limited and how can these limits be reflected in information foraging models? A thorough understanding of the nature and limits of human cost estimation would have important implications for the design of information seeking tools since it would give information on how best to convey cost information.

In this research models of scent and cost estimation were built on a static pool of data corresponding to an expert's past experience. This neglects the role recent experience might have on modulating estimates of scent and cost. For example, an initially high estimation of tag scent might be reduced upon discovery that the first few paragraph cells are irrelevant. Further work is needed to assess whether and how this dynamic modulation of scent and cost estimates occurs in humans and on how to represent this in foraging models.

Foraging models currently use an overly restrictive definition of cost. In addition to time, other quantifiable factors such as physical effort (measured as amount of movement, actions, energy) or mental workload (measured as working memory load using NGOMLS, and number of cognitive operations using CPM-GOMS). It has been argued that this broad array of user costs can be objectively measured using physiological measures (Wilson & Sasse 2000). It might thus be possible to validate sophisticated models and quantifications of user costs using such physiological measures.

Finally, information foraging theory needs to be extended to account for behaviour that conforms to the 'satisficing rule'. The satisficing rule states that a search will be terminated when a 'good enough' item has been found (see chapter 6, section 2.1). In biology, it is generally safe to assume that foraging animals never get to eat as much food as they would like. However in the case of information seeking this is not the case, searches terminate when the information need has been met, hence the 'satisficing rule'. Perhaps this could be integrated into information foraging theory via the notion that, due to overlaps between relevant items, there is a declining return on investment. Thus there comes a point where the cost of obtaining additional relevant material outweighs the additional benefit it will bring.

References

Adler, A., Gujar, A., Harrison, B. L., O'Hara, K., & Sellen, A. (1998), "A diary study of work-related reading: design implications for digital reading devices", in *Conf. Proc. On Human factors in computing systems*, pp. 241-248

Ahonen, H., Heikkinen, B., Heinonen, O., & Kilpeläinen, P. (1997), "Assembling document form digital libraries"; in *Lecture Notes in Computer Science*, 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA '97) (eds.), Springer Verlag, Toulouse, France.

American Library Association (1996), "Marbi, the USMARC Formats: Background and Principles", *Library of congress network development and MARC standards Office* http://lcweb.loc.gov/marc

Anderson, J. R. (1990), "The adaptive character of thought" Lawrence Erlbaum Associates, Hillsdale, NJ.

Anderson, J. R. (1993), "Rules of the mind" Lawrence Erlbaum, Hillsdale, NJ.

Attfield, S. & Dowell, J. (2003), "A study of information seeking and use by newspaper journalists", in *Journal of documentation*, vol. 59, no. 2,

Baeza-Yates, R. & Ribeiro-Neto, B. (1999), "Modern information retrieval" Addison Wesley, New York.

Beagle, D. (1999), "Visualisation of metadata", in *Information technology and libraries*, vol. 18, no. 4, pp. 193-199

Becker, R. A. & Cleveland, W. S. (1987), "Brushing Scatterplots", in *Technometrics*, vol. 29, no. 2, pp. 127-142

Bederson, B. B. (2000), "Fisheye Menus", in *Proceedings of ACM Conference on User Interface Software and Technology (UIST 2000)*, ACM Press, pp. 217-226

Benjamins, R. V., Fensel, D., Decker, S., & Perez, A. G. (1999), "(KA)2: Building ontologies for the internet: A mid-term report", in *International Journal of Human-Computer Studies*, vol. 51, pp. 687-712

Berners-Lee, T., Hendler, J., & Lassila, O. (2001), "The semantic web", in *Scientific american*, vol. 284, no. 5, pp. 34-43

Beyer, H. R. & Holtzblatt, K. (1995), "Apprenticing with the customer", in *Communications of the ACM*, vol. 38, no. 5, pp. 45-52

Bhavnani, S. K. & John, B. E. (2001), "The strategic use of complex computer systems"; chapter 5 in *Human-comuter interaction in the new millenium*, Carrol J.M. (eds.), ACM press, Addison Wesley, pp. 97-124.

Blair, D. C. & Maron, M. E. (1985), "An evaluation of retrieval effectiveness for a full-text document-retrieval system", in *Communications of the ACM*, vol. 28, no. 3, pp. 289-299

Boguraev, B., Kennedy, C., Bellamy, R., Brawer, S., Wong, Y. Y., & Swartz, J. (1998), "Dynamic Presentation of Document Content for Rapid On-Line Skimming", in *AAAI Spring 1998 Symposium on Intelligent Text Summarization*, pp. 118-128

Bolter, J. D. (1991), "Writing spaces: the computer, hypertext and the history of writing" Erlbaum, Hillsdale, N.J.

Botsch, V. & Kunz, C. (1996), "Visualization and navigation of networked information spaces: the matrix browser", in *Proceedings WebNet '96*, San Francisco, CA, USA, pp. 101-104

Bovair, S., Kieras, D. E., & Polson, P. G. (1990), "The acquisition and performance of text-editing skill: a cognitive complexity analysis", in *Human computer interaction*, vol. 5, pp. 1-45

Brand, M. (1997), "Discovering and visualizing narrative themes", in *Proceedings of Visual 97*,

Brown, M. H. & Weihl, W. E. (1996), "Zippers: A Focus+Context Display of Web Pages", in *Electronic proceedings of Webnet 96, AACE*,

Bryce, A. (1997), "Information Needs: A Person-in-Situation Approach", in *Information Seeking in Context*, P. Vakkari, R. Salvolainen, & B. Dervin, eds., Taylor Graham, London, pp. 111-122

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2000), "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices", in *Proceedings of the Tenth International World-Wide Web Conference*, http://www10.org/cdrom/papers/594/

Byrd, D. (1999), "A Scrollbar-based Visualization for Document Navigation", in *Proceedings of Digital Libraries 99 Conference*, ACM Press, New York, pp. 22-29

Byrne, M. D. (2002), "Cognitive Architecture"; chapter 5 in *Human-Computer Interaction Handbook.*, J. Jacko & A. Sears (eds.), Erlbaum, Hillsdale, NJ.

Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. (1999), "The tangled web we wove: A taskonomy of WWW use", in *Human Factors in Computing Systems: Proceedings of CHI 99*, ACM Press, pp. 544-551

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999), "Information visualization: Using vision to think" Morgan Kaufmann, San Francisco.

Card, S. K., Moran, T., & Newell, A. (1983), "The psychology of human computer interaction" Erlbaum, Hillsdale, New Jersey.

Card, S. K., Pirolli, P., & Mackinlay, J. D. (1994), "The cost-of-knowledge characteristic function: display evaluation for direct-walk dynamic information visualizations", in *Proceedings of CHI'94*, ACM Press, pp. 238-244

Card, S. K., Robertson, G., & Mackinlay, J. D. (1991), "The information visualizer: an information workspace", in *ACM Conference on human factors in computing systems*, ACM Press, New Orleans, USA, pp. 181-188

Card, S. K., Robertson, G., & York, W. (1996), "The WebBook and the Web Forager: An information workspace for the World-Wide Web", in *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, ACM Press, pp. 111-117

Carey, M., Kriwaczek, F., & Ruger, S. M. (2000), "A Visualization Interface for Document Searching and Browsing", in *Proc of NPIVM 2000*, http://www.doc.ic.ac.uk/~srueger/www-pub/vidsb-npivm2000.pdf

Carey, M. S., Stammers, R. B., & Astley, J. A. (1989), "Human-Computer Interaction Design: the Potential and Pitfalls of Hierarchical Task Analysis"; in *Task Analysis for Human-Computer Interaction*, D. Diaper (eds.), Halsted Press, New York, pp. 56-73.

Carroll, J. M. (2000), "Five reasons for scenario based design", in *Interacting with Computers*, vol. 13, pp. 43-60

Chen, C. & Czerwinski, M. (2000), "Empirical evaluation of information visualisations: an introduction", in *Int.J.Human-Computer Studies*, vol. 53, pp. 631-635

Chen, C. & Paul, R. J. (2001), "Visualizing a knowledge domain's intellectual strucutre", in *IEEE computer*, vol. 34, pp. 65-71

Cockburn, A. & McKenzie, B. (2000), "An evaluation of cone trees", in *People and Computers XV (Proceedings of the 2000 British Computer Society Conference on HumanComputer Interaction.)*, Springer-Verlag, University of Sunderland,

Collins, A. M. Q. M. R. (1969), "Retrieval from semantic memory", in *Journal of Verbal Learning and Verbal Behavior*, vol. 8, pp. 240-247

Crestani, F. & Ntioudis, S. P. (2001), "User centered evaluation of an automatically constructed Hyper-TextBook", in *Journal of educational multimedia and hypermedia*, vol. 11, no. 1, pp. 3-19

Cutting, D. R., Karger, D. R., & Pedersen, J. O. (1992), "Scatter/gather: A cluster-based approach to browsing large document collections", in *Proceedings of the SIGIR '92*, ACM, New York, pp. 318-329

de Bruijn, O. & Spence, R. (2000), "Rapid Serial Visual Presentation: A Space-Time Trade-Off in Information Presentation", in *Proceedings of AVI2000*, ACM Press, Palermo, Italy, pp. 189-192

De Carvalho Moura, A. M., Campos, M. L., & Barreto, C. M. (1998), "A survey of metadata for describing and retrieving internet resources", in *World Wide Web*, vol. 1, pp. 221-240

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T. K., & Harshman, R. (1990), "Indexing by Latent Semantic Analysis", in *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407

Ding, W. & Marchionini, G. (1998), "Overviews and previews for multimedia instructional resources", in *Proceedings of ASIS'98*, http://www.learn.umd.edu/reports/asis/98

Dourish, P., Lamping, J., & Rodden, T. (1999), "Building bridges: customisation and mutual intelligibility in shared category management", in *Proceedings of the international ACM SIGGROUP conference on Supporting group work*, Phoeniz, AZ, pp. 11-20

Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhardt, J., & Lochbaum, C. C. (1989), "Formative design evaluation of SuperBook", in *ACM Transactions on Information Systems*, vol. 7, no. 1, pp. 30-57

Ehn, P. (1988), "Work-oriented design of computer artifacts" Arbetslivscentrum, Stockholm.

Eick, S. G., Steffen, J. L., & Sumner, E. E. (1992), "SeeSoft – A tool for visualizing line oriented software statistics", in *IEEE Transactions on Software Engineering*, vol. 18, pp. 957-968

Erdelez, S. (1996), "Information encountering: a conceptual framework for accidental information discovery", in *Information seeking in context: Proceedings of an international conference on research in information needs, seeking and use in different contexts*, Taylor Graham, Los Angeles, CA, pp. 412-421

Fairchild, K. M., Poltrock, S. E., & Furnas, G. (1988), "SemNet: Three-Dimensional Representations of Large Knowledge Bases"; in *Cognitive science and its applications for human-computer interaction*, G. Raymonde (eds.), Lawrence Erlbaum, Hillsdale, NJ, pp. 201-233.

Fox, M. S. (1992), "The TOVE Project: A Common-sense Model of the Enterprise"; in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, F. Belli & F. J. Radermacher (eds.), Springer-Verlag, Berlin, pp. 25-34.

Furnas, G. (1981), "The FISHEYE view: a new look at strucutred files"; chapter 4.1 in *Information visualization: Using vision to think*, pp. 312-330.

Furnas, G. (1986), "Generalised fisheye views", in *CHI'86*, ACM Press, pp. 16-23

Furnas, G., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987), "The vocabulary problem in human-system communication", in *Communications of the ACM*, vol. 30, no. 11, pp. 964-971

Gaines, B. & Shaw, M. L. G. (1993), "Knowledge Acquisition Tools based on Personal Construct Psychology", in *Knowledge Engineering Review*, vol. 8, no. 1, pp. 49-85

Ginsburg, A., Marks, J., & Shieber, S. (1997), "A view for postscript documents", in *Proceedings of the 9th annual ACM symposium on User interface software and technology*, pp. 31-32

Gong, R. (1993), Validating and refining the GOMS model methodology for software user interface design and evaluation, Phd. Thesis, University of Michigan.

Graham, J. (1999), "The reader's helper: a personalised document reading environment", in *CHI' 99*, ACM Press, Pittsburgh PA, USA,

Gray, W. D. & Boehm-Davis, A. (2000), "Milliseconds matter: an introduction to microstrategies and their use in describing and predicting interactive behaviour", in *Journal of Experiment Psychology: Applied*, vol. *6*, no. 4, pp. 322-335

Gray, W. D., John, B. E., & Atwood, M. E. (1993), "Project Ernestine: a validation of GOMS for prediction and explanation of real-world tasks", in *Human computer interaction*, vol. 8, no. 3, pp. 237-309

Gruninger, M. & Fox, M. S. (1995), "The logic of enterprise modeling"; in *Reengineering the enterprise*, J. Brown & D. O'Sullivan (eds.), Chapman and Hall, pp. 83-98.

Halasz, F. & Schwartz, M. (1994), "The dexter hypertext reference model", in *Communications of the ACM*, vol. 37, no. 2, pp. 30-39

Harman, D. (1993), "Overview of the first text retrieval conference", in *Proceedings of the 16th annual international ACM/SIGIR conference*, pittsburgh, PA, pp. 36-38

Harman, D., Schauble, P., & Smeaton, A. (1997), "Document Retrieval"; chapter 7.2 in *Survey of the state of the art in human language technology*, R. Cole (eds.), Cambridge University Press, Cambridge.

Harris, S. L. & Ingram, J. H. (1995), "Real information, virtual documents", in *Proceedings of the 13th annual international conference on Systems documentation: emerging from chaos: solutions for the growing complexity of our jobs*, Savannah, Georgia, United States, pp. 71-76

Hayes, J. R. (1996), "A new framework for understanding cognition and affect in writing"; in *The science of writing : theories, methods, individual differences and applications*, S. Ransdell & M. Levy (eds.), Erlbaum, Mahwah, N.J.

Healey, C. (1996), "Choosing effective colours for data visualization", in *Proceedings of the Conference on Visualization '96*, pp. 263-270

Healey, C., Booth, K., & Enns, J. (1996), "High-speed visual estimation using preattentive processing", in *ACM Transactions on Computer-Human interactions*, vol. 3, no. 2, pp. 107-135

Hearst, M. (1995), "TileBars: Visualization of term distribution information in full text information access", in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, Denver, CO, pp. 59-66

Hearst, M. (1999), "User Interfaces and Visualization"; chapter 10 in *Modern information retrieval*, R. Baeza-Yates & B. Ribeiro-Neto (eds.), Addison-Wesley.

Hearst, M. & Karadi, C. (1997), "Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy", in *Proceedings of the 20th ACM/SIGIR conference*,

Hearst, M. & Plaunt, C. (1993), "Subtopic structuring for full-length document access", in *Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval*, ACM Press, p. 59

Holmquist, L. E. (1997), "Focus+Context visualization with flip zooming and the zoom browser", in *ACM SIGCHI*, ACM Press, pp. 263-264

Hornbæk, K. & Frøkjær, E. (1999), "Do Thematic Maps improve information retrieval?", in *Proceedings of Human-Computer Interaction—INTERACT '99*, IOS Press, Amsterdam, The Netherlands, pp. 179-186

Hornbæk, K. & Frøkjær, E. (2001), "Reading Electronic Documents: The Usability of Linear, Fisheye, and Overview+Detail Interfaces", in *Proceedings of CHI'2001*, pp. 293-300

Howell, D. C. (1997), "Statistical methods for psychology", 4th edn., Wadsworth publishing company.

Huc, C., Levoir, T., & Nonon-Latapie, M. (1997), "Metadata: models and conceptual limits", in *Proceedings of the second IEEE metadata conference*, Silver Spring, Maryland,

Hyperionics (2002), "HyperCam", *Hyperionics Technology, LLC* http://www.hypercam.com/

IDC (11-1-2001), "Portals Follow Intranets", *IDC* http://www.idc.com/getdoc.jhtml?containerId=ebt20010111

Ingwersen, P. (1996), "Cognitive perspectives of information retrieval interaction elements of a cognitive IR theory", in *Journal of documentation*, vol. 52, no. 1, pp. 3-50

Janney,K. and Sledge, J. (1995), "A user model for CIMI Z39.50 application profile" *CHIO* www.cimi.org/old_site/documents/Z3950_app_profile_0995.html

Joachims, T. (1999), "Transductive Inference for Text Classification using Support Vector Machines", in *International Conference on Machine Learning*, pp. 200-209

John, B. E. & Kieras, D. E. (1996), "The GOMS family of user interaface analysis techniques: comparison and contrast", in *ACM Transactions on Computer-Human interactions*, vol. 3, no. 4, pp. 320-351

John, B. E. & Newell, A. (1989), "Cumulating the science of HCI: From S-R compatibility to Transcription typing", in *CHI proceedings*, ACM Press, pp. 109-114

John, B. E. & Newell, A. (1990), "Towards an engineering model of stimulus-response compatibility"; R. Proctor & T. Reeve (eds.), Elsevier, New York, pp. 427-479.

Johnson, J. D. (2002), "On contexts of information seeking", in *Information processing and management* – In press

Johnson, P. (1992), "Human computer interaction: psychology, task analysis and software engineering" McGraw Hill.

Jokela, S., Turpeinen, M., & Sulonen, R. (2000), "Ontology development for flexible

content", in *Proceedings of the 33rd Hawaii International Conference on System Sciences*, IEEE Computer Society Press, pp. 6056-6066

Kando, N. (2001), "NTCIR Workshop: Japanese- and Chinese-English cross-lingual information retrieval and multi-grade relevance judgments", in *Lecture Notes in Computer Science; 2069*, Springer, Lisbon, Portugal, pp. 24-33

Kieras, D. E. (1994), A guide to GOMS task analysis, *Technical Report,* University of Michigan, www.citeseer.nj.nec.com/kieras94guide.html

Kieras, D. E. (1996), "A guide to GOMS model usability evaluation using NGOMSL"; chapter 31 in *Handbook of Human-Computer Interaction*, M. G. Helander, T. K. Landauer, & P. V. Prabhu (eds.), North-Holland, pp. 733-766.

Kieras, D. E. & John, B. E. (1996), "Using GOMS for user interface design and evaluation: which technique?", in *ACM Trans. Comput.-Hum. Interact.*, ACM, pp. 287-319

Korn, F. & Shneiderman, B. (1995), Navigating Terminology Hierarchies to Access a Digital Library of Medical Images, *Dept. of Computer Science Technical Report,* University of Maryland,

Kuhlthau, C. C. (1993), "A principle of uncertainty for information seeking", in *Journal of documentation*, vol. 49, no. 4, pp. 339-355

Kuhlthau, C. C. & Tama, S. L. (2001), "Information search process of lawyers: a call for 'just for me' information services", in *Journal of documentation*, vol. 51, no. 1, pp. 25-43

Kumar, H. P., Plaisant, C., & Shneiderman, B. (1997), "Browsing Hierarchical Data with MultiLevel Dynamic Queries and Pruning", in *International Journal of Human-Computer Studies*, vol. 46, no. 1, pp. 103-124

Lamping, J. & Rao, R. (1996), "Visualizing large trees using the hyperbolic browser", in *Proceedings of ACM UIST '94* , pp. 13-24

Landoni, M., Crestani, F., & Melucci, M. (2000), "The Visual Book and the Hyper-TextBook: Two Electronic Books One Lesson?", in *RIAO'2000 Proceedings*, Paris, France, pp. 247-265

LCSH (2002), "Library of Congress Subject Headings", *Library of congress* http://www.loc.gov

Levy, D. M. (1997), "I read the news today, oh boy: reading and attention in digital libraries", in *Proceedings of the second ACM international conference on Digital libraries*, pp. 202-211

Mackinlay, J. D. & Robertson, G. (1993), "The Document Lens", in *Proceedings of the ACM User Interface and Software Technology conference (UIST'93)*, ACM Press, pp. 101-108

Malone, T. W. (1983), "How do people organize their desks? Implications for the design of office systems", in *ACM Transactions on office information systems*, vol. 1, no. 1, pp. 99-112

Marchionini, G. & Shneiderman, B. (1988), "Finding Facts vs. Browsing Knowledge in Hypertext Systems", in *IEEE computer*, vol. 21, no. 1, pp. 70-80

McCracken, G. D. (1988), "The long interview" Sage Publications, Newbury Park; London.

MESH (2002), "Medical Subject Headings", *National Library of Medicine, Bethesda, MD.* http://www.nlm.nih.gov

Miller, C. S. & Remington, R. W. (2001), "Modeling an opportunistic strategy for information navigation", in *Proceedings of the Twenty-Third Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, Edinburgh, http://www.hcrc.ed.ac.uk/cogsci2001/pdf-files/0639.pdf,

Miller, N. E., Wong, P. C., Brewster, M., & Foote, H. (1998), "TOPIC ISLANDS™ - a wavelet-based text visualisation system", in *IEEE Visualization '98*,

Morse, E. & Lewis, M. (2000), "Evaluating visualizations: using a taxonomic guide ", in *International Journal of Human-Computer Studies*, vol. 53, no. 5, pp. 637-662

Motta, E., Buckingham Shum, S., & Domingue, J. (2002), "Ontology-driven document enchrichement: principles, tools and applications", in *Int.J.Human-Computer Studies*, vol. 52, no. 5, pp. 1071-1109

Mukherjea, S., Foley, J. D., & Hudson, S. E. (1995), "Visualizing complex hypermedia networks through multiple hierarchical views", in *Proceeding of CHI'95, ACM Conference on human factors in computing systems*, pp. 331-337

Mullet, K. & Sano, D. (1995), "Designing visual interfaces: communication oriented techniques." Sunsoft Press, Moutain View; CA.

Mynatt, B. T., Leventhal, L. M., Instone, K., Farhat, J., & Rohlman, D. S. (1992), "Hypertext or book: Which is better for answering questions", in *Proceedings of ACM CHI 1992*, pp. 11-19

Neave, H. R. & Worthington, P. L. (1988), "Distribution free tests" Unwin Hyman, London.

Newman, W. (1997), "Better or just different? On the benefits of designing interactive systems in terms of critical parameters", in *Proceeding of DIS '97, Designing interactive systems*, Amsterdam, The Netherlands, pp. 239-245

Nicholas, D. & Martin, H. (1993), "Should journalist search themselves? (And what happens when they do)", in *Online Information 93*, pp. 227-234

Norman, K. L. (1991), "The psychology of menu selection: designing cognitive control at the human/computer interface" Ablex publishing corporation, Norwood, New Jersey.

North, C. & Shneiderman, B. (2000), "Snap-together visualization: can users construct and operate coordinated visualizations?", in *Int.J.Human-Computer Studies*, vol. 53, pp. 715-739

O'Day, V. & Jeffries, R. (1993), "Orienteering in an information landscape: how information seekers get from here to there", in *InterCHI'93*, pp. 438-445

O'Hara, K. & Sellen, A. (1997), "A comparison of reading paper and on-line documents", in *Proceedings of CHI'97 Conference on Human Factors in Computing Systems*, ACM Press, pp. 335-342

O'Hara, K., Taylor, R. S., Newman, W., & Sellen, A. (2002), "Understanding the materiality of writing from multiple sources", in *Int.J.Human-Computer Studies*, vol. 56, pp. 269-305

Ogden, W., Davis, M., & Rice, S. (1999), "Document thumbnail visualizations for rapid relevance judgements: When do they pay off?", in *The Seventh Text REtrieval Conference (TREC-7)*, E. M. Voorhees & D. Harman, eds., Gaithersburg, MD, USA., pp. 528-534

Olson, J. R. & Olson, G. M. (1990), "The growth of cognitive modeling in human-computer interaction since GOMS", in *Human computer interaction*, vol. 5, pp. 221-265

Pepper, S. (1998), "Navigating haystacks and discovering needles", in *Markup Languages: Theory & Practice*, vol. 1, no. 4, pp. 47-74

PICS (1996), "PICS: Internet Access Controls Without Censorship", in *Communications of the ACM*, vol. 39, no. 10, pp. 87-93

Pirolli, P. (1997), "Computational Models of Information Scent-Following in a very Large Browsable Text Collection", in *Proceedings of theConference on Human Factors in Computing Systems,CHI '97*, Atlanta, GA, pp. 3-10

Pirolli, P. (1998), "Exploring Browser Design Trade-Offs Using a Dynamical Model of Optimal Information Foraging", in *Proceedings of theConference on Human Factors in Computing Systems, CHI '98*, Los Angeles, pp. 33-40

Pirolli, P. & Card, S. K. (1998), "Information Foraging Models of Browsers for Very Large Document Spaces", in *AVI '98*, L'Aquilla, Italy, pp. 83-93

Pirolli, P. & Card, S. K. (1999), "Information Foraging", in *Psychological Review*, vol. 106, no. 4, pp. 643-675

Pirolli, P., Fu, W., Reeder R., & Card, S. K. (2002), "A user-tracing architecture for modeling interaction with the World Wide Web", in *Advanced Visual Interfaces (AVI 2002)*, Trento, Italy.,

Pirolli, P. & Rao, R. (1996), "Table lens a tool for making sense of data", in *Proceedings of AVI '96*, ACM Press, pp. 67-80

Pollitt, A. S. (1997), "Interactive Information Retrieval based on Faceted Classification using Views, Knowledge organization for Information Retrieval", in *Proceedings of the 6th International Study Conference on Classification*, University College, London, pp. 51-56

Prettyman, R. M. (1996) Conversion to SGML: The Pain, The Gain. <TAG> *9*[*9*], 6-9.

Reder, L. M. (1979), "The role of elaborations in memory for prose", in *Cognitive psychology*, vol. 11, pp. 221-234

Risden, K., Czerwinski, M., Munzner, T., & Coook, D. (2000), "An Initial Examination of Ease of Use for 2D and 3D Information Visualizations of Web Content", in *Int.J.Human-Computer Studies*, vol. 53, pp. 695-714

Ritter, F. E. & Young, R. M. (2001), "Embodied models as simulated users: introduction to this special issue on using cognitive models to improve interface design", in *Int.J.Human-Computer Studies*, vol. 55, no. 1, pp. 1-14

Robertson, G., Cameron, K., Czerwinski, M., & Robbins, D. (2002), "Polyarchy Visualization: Visualizing multiple intersecting hierarchies", in *ACM SIGCHI 2002*, ACM Press, pp. 423-430

Robertson, G., Mackinlay, J. D., & Card, S. K. (1991), "Cone Trees: Animated 3D visualizations of hierarchical information", in *Proceedings of CHI'91*, ACM Press, pp. 189-194

Rosch, E. (1978), "Principles of categorization"; in *Cognition and categorization*, E. Rosch & B. B. Lloyd (eds.), Erlbaum, Hillsdale, N. J., pp. 27-48.

Rubin, J. (1994), "Handbook of usability testing" John Wiley & Sons, New York.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993), "The cost structure of sensemaking", in *Conference proceedings on Human factors in computing systems*, ACM Press, pp. 269-276

Saracevic, T. (1995), "Evaluation of evaluation in information retrieval", in *SIGIR'95*, ACM Press, Seattle, CA, USA, pp. 138-146

Schilit, B. N., Price, M. N., & Golovchinsky, G. (1998), "Digital library information appliances", in *Digital libraries 98*, Pittsburg PA, USA, pp. 217-226

Schuldt,R.L. (1994), "Using a universal data element framework for enterprise integration", *Northern Virginia CALS Interest Group* http://www.llnl.gov/liv_comp/metadata/other-efforts/cals-briefing.ps

Sellen, A. (1996) Affordances of paper: phase I report. *Internal Report* Rank Xerox Research Laboratory, Cambridge, UK.

Sellen, A. & Harper, R. (1997), "Paper as an analytic resource for the design of new technologies", in *Conference proceedings on Human factors in computing systems*, pp. 319-326

Shneiderman, B. (1994), "Dynamic Queries for Visual Information Seeking", in *IEEE software*, vol. 11, no. 6, pp. 70-77

Shneiderman, B. (1996), "The eyes have it: a task by data type taxonomy for information visualisations", in *IEEE Workshop on Visual Languages*, pp. 336-343

Shneiderman, B. (1998), "Designing the User Interface: Strategies for Effective Human-Computer Interaction", 3rd edn., Addison-Wesley, Reading, MA.

Small, D. L. (1999), Rethinking the book, Ph.d. Thesis, MIT media lab.

Sparck Jones, K. (1995), "Reflections on TREC", in *Information processing and management*, vol. 31, no. 3, pp. 291-314

Spence, R. & Apperly, M. (1982), "Data base navigation: an office enviroment for the professional", in *Behaviour and Information Technology*, vol. 1, no. 1, pp. 43-54

Stephens, D. W. & Krebs, J. R. (1986), "Foraging Theory." Princeton University Press, Princeton, New Jersey.

Suh, B., Woodruff, A., Rosenholtz, R., & Glass, A. (2002), "Popout Prism: Adding Perceptual Principles to Overview+Detail Document Interfaces", in *Proc. CHI 2002*, ACM Press, Minneapolis, pp. 251-258

Sutcliffe, A. G., Ennis, M., & Hu, J. (2000), "Evaluating the effectiveness of visual user interfaces for information retrieval ", in *International Journal of Human-Computer Studies,* vol. 53, no. 5, pp. 741-763

Swan, R. C. & Allan, J. (1998), "Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems", in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 173-181

Taylor, R. S. (1968), "Question-negotiation and information seeking in libraries", in *College and research libraries*, vol. 29, pp. 178-194

TEI (2002), "Guidelines for electronic text encoding and interchange", *TEI,Text encoding initiative consortium* http://www.tei-c.org/

Tufte, E. R. (1990), "Envisioning information" Graphics Press, Cheshire, Connecticut.

Uschold, M. (1996), "Building ontologies: Towards a unified methodology", in *Proceedings of Expert Systems '96, 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, Specialist Group on Expert Systems* edn, Springer, Cambridge,

Uschold, M. & Gruninger, M. (1996) Ontologies: Principles, methods and applications. Knowledge Engineering Review, vol. 11, no. 2, 93-155.

Uschold, M., King, M., Moralee, S., & Zorgios, Y. (1998), "The Enterprise Ontology", in *The Knowledge Engineering Review*, vol. 13, no. 1, pp. 31-89

Vakkari, P. (2001), "A theory of task-based information retrieval", in *Journal of documentation*, vol. 57, no. 1, pp. 44-60

Van Der Wege, M. M. & Pirolli, P. (2000), "The effect of information scent on searching information visualizations of large tree structures", in *Advanced Visual Interfaces, AVI 2000*, ACM Press, Palermo, Italy, pp. 161-172

Veerasamy, A. & Belkin, N. J. (1996), "Evaluation of a tool for visualization of information retrieval results", in *SIGIR'96* , ACM Press, Zurich, Switzerland, pp. 85-92

Veerasamy, A. & Heikes, R. (1997), "Effectiveness of a Graphical Display of retrieval results", in *SIGIR 97*, ACM Press, Philadelphia, USA, pp. 236-245

Voorhees, E. M. (1998), "Variations in relevance judgements and the measurement of retrieval effectivness", in *SIGIR'98*, ACM Press, Melbourne, Australia, pp. 315-323

Voorhees, E. M. & Harman, D. (1998), "Overview of the seventh text retrieval Conference (TREC-7)", in *The Seventh Text Retrieval Conference*, Gaithersburg, MD, USA, pp. 1-23

Voorhees, E. M. & Harman, D. (2000), "Overview of the Eighth Text REtrieval Conference (TREC-8", in *Proceedings of the Eighth Text REtrieval Conference)*, Gaithersburg, Maryland, pp. 1-24

W3C (1998), "Recommendation: Extensible Markup Language (XML) 1.0", *W3C* http://www.w3.org/TR/1998/REC-xml-19980210

W3C (5-1-1999), "Resource Description Framework (RDF) Model and Syntax W3C Proposed Recommendation", *W3C* http://www.w3.org/TR/PR-rdf-syntax/

Ware, C. (2000), "Information Visualization: Perception for Design" Morgan Kaufman, San Fancisco.

Weibel,S., Grodby, J., and Miller, E. (1995), "OCLC/NCSA Metadata Workshop Report", *Dublin, USA* http://www.ifla.org/documents/libraries/cataloging/oclcmeta.htm

Wilson, G. & Sasse, M. A. (2000), "Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality", in *Proceedings of HCI 2000: People and Computers XIV - Usability or Else!*, S. McDonald, Waern Y., & Cockton G., eds., Sunderland, UK, pp. 327-339

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995), "Visualising the non-visual: spatial analysis and interaction with information from text documents", in *Proceeding of IEEE information visualisation'95*, pp. 51-58

Wong, Y. Y. (1995), Temporal typography. Characterization of time-varying typographic forms, Master's Thesis, MIT Media Lab.

Woodruff, A., Gossweiler, R., Pitkow, J., Chi, E. H., & Card, S. K. (2000), "Enhancing a Digital Book with a Reading Recommender", in *Proc. CHI 2000*, ACM Press, The Hague, The Netherlands, pp. 153-160

Appendix 1 – Results and discussion for the overview and sorting task in the user study

This Appendix describes and discusses results from the user study reported in chapter 5 and section 1.

The results for the overview tasks, shown in Table 27, shows that summaries written based solely on inspection of GridVis are very different and inferior to those written using the full document text. Participants C and S wrote totally new and different summaries, despite having the previous one in view. Participants D and R used some or all of their initial propositions, correcting about a quarter of them, and adding many new ones. The propositions made in the GridVis based summaries were qualitatively different to those made with the benefit of the document text. They tended to be simple conjunctions of the objects deemed relevant (see Table 28). Participants D and R also produced more complex propositions, which, interestingly, needed to be corrected when reused in the summary based on the document. These results show that using GridVis alone only allows for a limited understanding of the document content.

	Reuse of objects reused/total from doc	Reuse of propositions (refined + same)/no of prop from doc	New propositions from doc added/no of props from doc	Propositions omitted from doc summary omitted/no of props from GridVis	Incorrect propositions from GridVis omitted/no of objects from GridVis
Participant C	0.07	0.00	1.00	1.00	0.00
Participant S	0.29	0.00	1.00	1.00	0.00
Participant D	0.10	0.25	0.88	0.50	0.25
Participant R	0.43	0.44	0.56	0.14	0.29
Average	**0.22**	**0.17**	**0.86**	**0.66**	**0.13**

Table 27 Results for overview task

	Propositions	
Using GridVis	Proposal is something to do with stock allocation, depots and MORs	MOR stock allocation rules adjustment Simulation with results
Using the full document text	Modelling has shown that MOR rules allow the depot to operate within capacity.	Use of simulation to see the effects of scaling down MOR

Table 28 Example of the difference between propositions created using GridVis and the full document text

In the sorting task, participants made few changes to their ratings upon seeing the full document text. On average they changed one out of the six ratings and the Spearman's Rank Correlation between the before and after ratings was 0.89. This suggests that the GridVis visualisation is a good for this kind of sorting task, as an inspection of the full document text.

There was a substantial difference between the ratings given by the author and the readers; on average four out of the six ratings were different. This error was not consistently reduced when the participants were given the full document text. In retrospect, these particular results were probably poor because only a single judge was used to create the benchmark ratings. The experimental evaluation (chapter 5, section 3) used several judges and found considerable variation between judges.

1 Discussion

GridVis does not give a rich overview of document content, and as such, it does not support browsing as defined by Adler et al (Adler et al. 1998). The summaries written based on the visualisation did not express the document's contents and were thus either completly rewritten, or substantially improved once the full document was made available. As such GridVis does not effectively allow the user to "obtain a general sense of [...] content and extent" p. 244 (op. cit.).

GridVis did seem to support the sorting task well if the performance is compared to that with access to the full document text. Users only felt the need to make small changes to the ratings once they had seen the full document.

Although sorting was well supported by GridVis, it was not clear that GridVis offered any intrinsic advantages over existing alternatives. A user of GridVis tags can pick out the tags that apply to large portions of the documents. These indicate which objects are being discussed which, evidently, is sufficient to do the sorting. However, sorting can currently also be done using devices normally incorporated into business documents such as the management summary and section titles. Since these also succinctly convey the main topics covered by the document, it is not clear that GridVis has the potential to offer a clear advantage for this task.

Appendix 2 – Expert judge ratings

Retrieval performance was measured by comparing the paragraph relevance ratings made by the participants to a benchmark set of ratings made by expert judges.

Each document had the relevance of each paragraph rated, for each question, by three judges. One of the judges was the document's author and the other two were participants in the experiment. The relevance of each paragraph was rated on a three point scale, as one of '3 - contains key information', '2 - contains relevant information', and '1 - contains related information' (see Table 29).

Rating	Rating description
3	Contains key information (i.e. necessary to answering the question).
2	Contains relevant information (i.e. would improve the quality of the answer but is not essential)
1	Contains related information (i.e. could marginally improve the quality of the answer but is not essential)

Table 29 The ratings scale given to judges and participants

Examination of these judge ratings made it clear that there was a great deal of variation between judges; over half the paragraphs rated as relevant were designated as such by only one judge. An analysis was carried out in order to characterise this variation and determine how best to create one set of ratings which captured multiple perspectives while excluding erroneous and highly idiosyncratic relevance judgements. Two of the four documents were used to do this analysis. The ratings were classified according to the interpretation of the question each rating implied. Also, the number of judges who rated a particular paragraph as relevant was counted; paragraphs which more than one judge rated as relevant, will be referred to below as having been 'corroborated'.

There tended to be at least two comprehensible interpretations of a question with one being more common than the other (48% of ratings versus 16%). The remaining 36% of ratings did not relate to any obvious interpretation of the question. This seems to be either due to misunderstandings of the questions or at least to very idiosyncratic understandings, since almost none (3%) of these were corroborated by other judges. Ideally, the aggregation scheme should retain ratings, which result from comprehensible interpretations of a question while omitting those resulting from incomprehensible interpretations.

One possible aggregation scheme would be to ignore all the uncorroborated ratings. This would remove 97% of the paragraphs which did not seem to relate to the question. Unfortunately, it would also remove all 27% of those classified as relating to the main theme. Another scheme would be to only ignore the uncorroborated paragraphs that were given a relevance rating of 1 or 2. This removes 90% of the paragraphs that don't seem to relate to the question and only 18% of those classified as relating to the main theme. An aggregation scheme, which ignored uncorroborated ratings of 1 and 2, but not rating of 3 would therefore be appropriate.

A simple aggregation scheme was designed to meet this requirement. For each paragraph, a mean was taken for the ratings from all 3 judges. The result was rounded up if it was above 1, and set to 0, if it was below 1. This has the effect of discounting uncorroborated 1's and 2's but reinforcing ratings on paragraphs where judges' ratings are corroborated.

A remaining problem with this aggregation was the distortion it could produce. Consider the possibility of a paragraph rated as 3 (i.e. 'contains key information') by two of the judges and as 0 by the third. The aggregation scheme would assign this

paragraph an overall rating of 2, even though it might have simply been misread by the third judge, and actually be extremely relevant to the question.

To deal with this problem, an adjudication meeting was used to decide the ratings of the most contentious paragraphs. In order to make best use of the limited time the most contentious paragraphs were selected using specific criteria. The paragraphs had to have a disparity in the relevance ratings of at least 2, and have been rated as 3, by at least one judge. These paragraphs are important since there is much disagreement on their rating but they might be very relevant.

The adjudication meeting was attended by all the judges. They were given copies of the questions, the rating scale and the documents which anonymously displayed the judges' ratings alongside each paragraph. Each of the selected contentious paragraphs were read, the judges discussed the answer and an agreement on the rating was reached. For some paragraphs, agreement proved elusive due to differences in interpretation of the question, the ratings scale, and the paragraph. For these paragraphs, the researcher was forced to act as an arbitrator, taking the discussion into consideration and deciding on the rating, before moving the meeting on to the next paragraph.

Appendix 3 - Configurations of tag selection strategies

This Appendix presents data on prevalence of difference tag selection strategies, from the experimental evaluation of GridVis reported in section 3 of chapter 5.

When the use of tag selection strategies in is examined closely, it is clear that there is much variation across participants. This makes the overview shown of tag selection strategies shown in section 3.3.1.2 of chapter 5 (pie chart of tag selection) somewhat misleading. It seems that different people use different configurations of strategies to select their tags. Participants with similar configurations of strategies were grouped together, as can be seen in the graphs below.

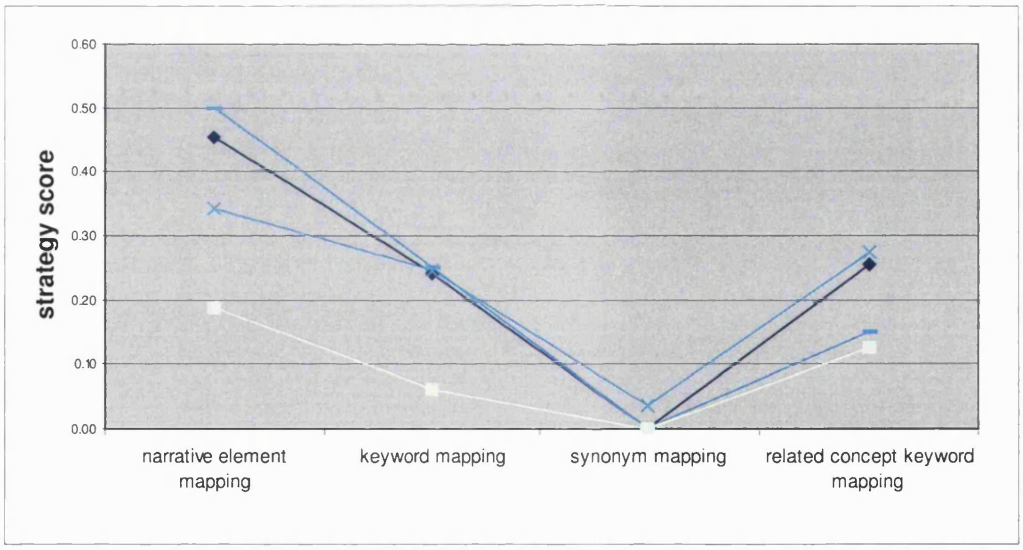


Figure 88 Scores for each tag selection strategy across four participants with a similar configuration of strategies, shown as percentage of total score.

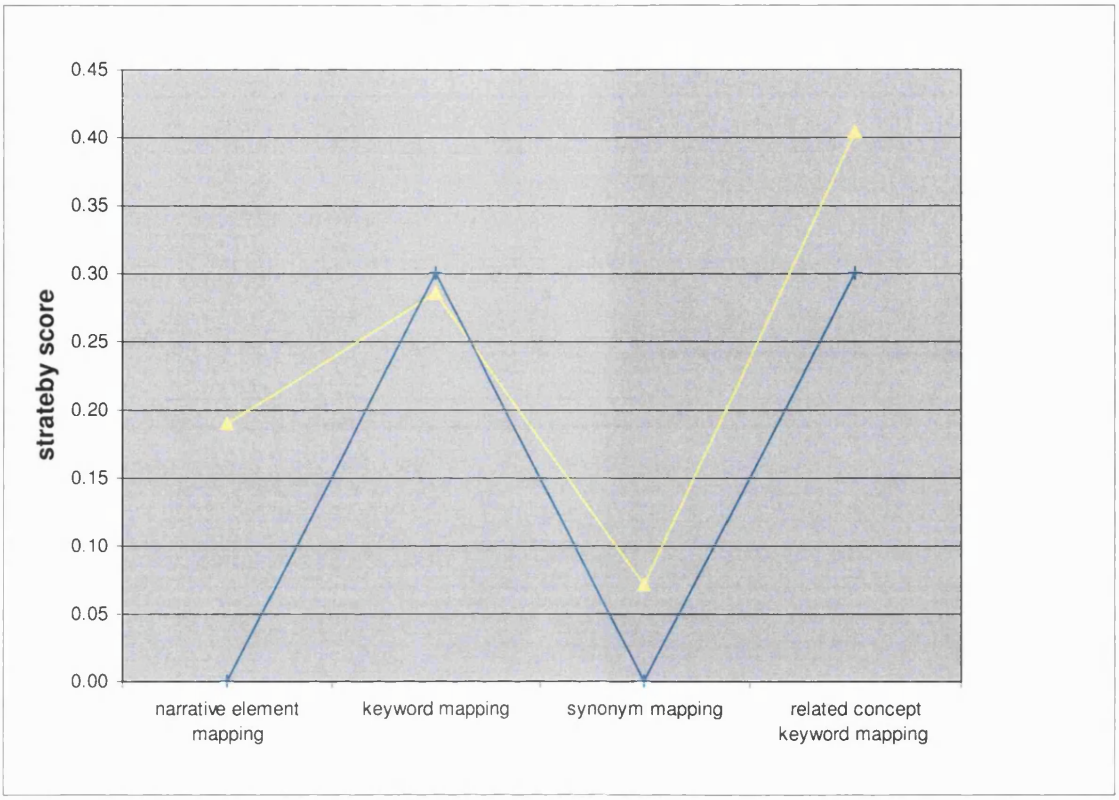


Figure 89 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score.

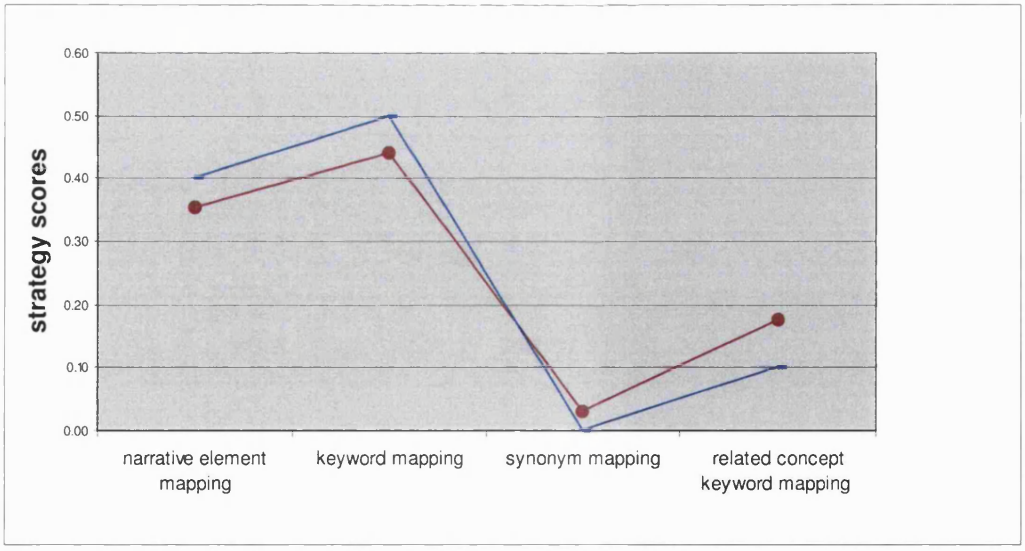


Figure 90 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score.

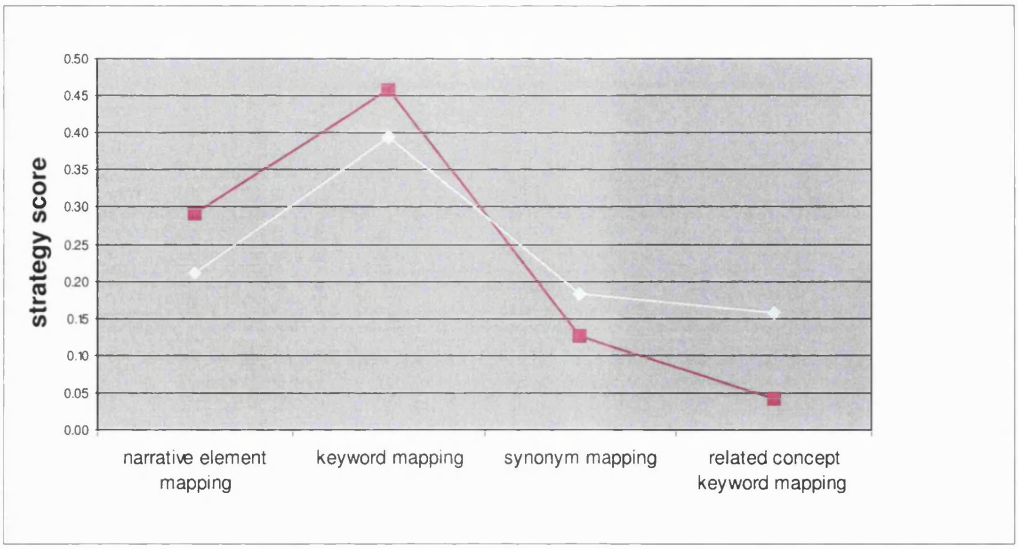


Figure 91 Scores for each tag selection strategy across two participants with a similar configuration of strategies, shown as percentage of total score.

These four graphs show the variation across participants in combinations of tag selection strategies used. In Figure 88 the 'narrative element mapping' strategy is primary, with the 'keyword mapping' and 'related concept keyword mapping' coming an equal second, and 'synonym mapping' hardly being used at all. On the other hand, in Figure 90 and Figure 91 the 'keyword mapping' strategy is primary while in Figure 89 it shares first place with the 'related concept keyword mapping' strategy. It is possible that the consistency in configuration of tag selection strategies within groups of participants seen here, especially in Figure 88, is indicative of a single, more complex, underlying strategy shared by each group. An alternative, and perhaps more plausible, explanation is that these differences are the result of variations in preferences and working styles.

Unfortunately, these results reveal little of import to this thesis. Although they show the prevalence of different tag selection strategies it is not clear how these relate to performance on goal-directed search, or to the design of GridVis.

Appendix 4 – GOMS models for tag and paragraph cell selection

This Appendix presents the full GOMS models for tag and paragraph selection. The operators invoked by these models are listed in chapter 6, sections 2.2.3.1.1 and 2.2.3.1.3. In section 4, this Appendix also includes the derivation of the get-from-task <cell-applicability> operator.

2 GOMS model for tag selection in the GridVis Interface

2.1 High level goals

Goal 0 – select the next most relevant tag

Goal 2 – decide on most relevant un-examined tag in <subsection>

Goal 3 – select <tag>

2.2 Methods

Method for goal: 0 select next most relevant tag

Step 1. Decide: if have just been investigating a tag from <section> then, goto Step 4

Step 2. Remember and home in on most relevant un-examined subsection

Step 3. Decide: if <section> is not worth inspecting then, return with goal accomplished

Step 4. Accomplish goal 2 <section>

Step 5. Decide: if <tag> is not worth inspecting then, goto Step 2

Step 6. Accomplish goal 3

Return with goal accomplished

Judgement call: Assuming that user is methodical to the point that they will stick to the subsection they last inspected, rather they starting all over again with a new section.

Method for goal:2 decide on most profitable un-examined tag in <section>

Step 1. Locate <section tag>

Step 2. Accomplish goal 2.1 (step2_cost = (goal2.1_cost*(n-k))/n)

Step 3. Decide: If there are no more tags in <section> then, goto Step 5

Step 4. goto Step 2

Step 5. Choose un-examined tag with highest profitability

Return with goal accomplished

Judgement call: This does not dissect the information scent sorting algorithm used. Because of this it is assumed that there is a fixed overhead for this (i.e. the decision time is independent of the number of tags). In practice this could be fairly accurate since there are often roughly the same number of tags in a section.

Judgement call: It is likely each time a section is revisited, a little more of the information scent data will be held in memory. It is not clear how this discounting of cost will occur. Therefore we assume the simplest case, where the cost of executing goal 2.1 decreases linearly reaching 0s at the last tag. So goal2.1_cost(n-k))/n were n, is the total number of tags, and k is the number of un-investigated tags. Cost is only discounted for this step and goal, since the other steps could not be speeded by information in memory accrued on last visit to the section.*

Judgement call: The section tag is included in the set of being considered. This is because the section tag sometimes has paragraph cells not assigned to any of its child tags. This means that the section tag can have a higher profitability than any one of its children.

Method for goal: 2.1 assess profitability of next unexamined tag

Step 1. Get-from-task <next tag>

Step 2. If tag already examined goto Step 1

Step 3. Compare to Query and Store info scent in WM[30]

Step 4. Get-from-task <number of cells on next tag><total width of cells on next tag> and Store in WM

Step 5. Calc profitability

Return with goal accomplished

Judgement call: How does the user check that they are exploring sections/tags that they haven't looked at yet? One could do it by reading down from the top of the list, checking against the contents of WM to find the next unexamined tag. Alternatively, information in WM about the location of the tag last selected could be used. Finally, a combination of both approaches could be used. Given the uncertainties, here it has been assumed that a single M-like operator can be can handle this.

Judgement call: It is likely each time a section is revisited, a little more of the information scent data will be held in memory. It is not clear how this discounting of cost will occur. Therefore the simplest case is assumed, where the cost of executing goal 2.1 decreases linearly reaching 0s at the last tag.

Method for goal: 3 Select <tag>

Step 1. locate and point to tag

Step 2. click tag & verify click

Return with goal accomplished

Judgement call: Locate and point operators lumped together, because Kieras suggests that this can be done for experts (Kieras 1996) and Gong (1993), found that locate,

---

[30] WM is used as an abbreviation for working memory

drops to 0 for experts. Click tag and verify click are put into the same step since Kieras suggests that for expert performace routinely followed steps with no interviening decisions can be put into one step (Kieras 1996).

3 GOMS model for Paragraph cell selection times

3.1 Left to right strategy

Method for goal: 1 select the next paragraph cell on <tag>

Step 1. visual search for first unselected cell on <tag> row &

point at <cell> and verify &

select <cell> and verify selection

Return with goal accomplished

Judgement call: All of these operators have been put into one step because Kieras recommends that for experts, often performed sequences of primitive operators which don't require any decisions can be put in one step (Kieras 1994).

Judgement call: In reality the mouse might begin to be moved towards the cell while the visual search is happening. Should CPMS-GOMS be used? – No, since visual search is too complicated to be modelled at that level, and it is not clear how to model the use of partial results to home the mouse to the right area.

3.2 Applicability only

Goal 1 – select the next highest unselected paragraph cell on <tag>

Method for goal: 1 select cell with highest information scent

Step 1. visual search for unselected cell on <tag> row with highest information scent

Step 2. point at <cell>, verify selection, and get-from-task, <cell-applicability>

Step 3. select <cell> and verify selection

Return with goal accomplished

Judgement call: The strategy used in this model assumes that the visual search is a single operation, which identifies the high scent cells and their location in one go. This accords with an introspective analysis of the analyst's strategy, in which the row is inspected and a gestalt is obtained which locates the regions with bright, medium bright, and dark cells.

Judgement call: In step 1 the time for the visual search has been set to 0, following Gong's suggestion that for experts the locate operator is set to 0 if followed by a point (Gong 1993). Since the locate operator is usually assigned a time of 1.2, it is essentially equivalent to the 'visual search for unselected cell on <tag> row with highest information scent' operator (Kieras & John 1996).

3.3 Co-occurrence only

Here only the scent of the tags co-occuring with the cell under question are considered.

Method for goal: 1 selects the unselected paragraphs on <tag> with the highest profitability

Step 1. visual search for first unselected cell on <tag> row

Step 2. point at <cell> cell verify, accomplish goal 4

Step 3. Decide: if visual search for next unselected cell on <tag> row finds a cell then, goto step 2

(total_cost_so_far = (cost_so_far *(n-k))/n)

Step 4. accomplish goal 3

Return with goal accomplished

Judgement call: It is likely each time a section is revisited, a little more of the information scent data will be held in memory. It is not clear how this discounting of cost will occur. Therefore we assume the simplest case, where the cost of executing goal 2.1 decreases linearly reaching 0s at the last tag. So goal2.1_cost(n-k))/n were n, is the total number of tags, and k is the number of un-investigated tags. Cost is only discounted for this step and goal, since the other steps could not be speeded by information in memory accrued on last visit to the section.*

Method for goal: 4 evaluate information scent of co-occurring tags

Step 1. visual search for <highlighted leaf tags>

Step 2. get-from-task <highlighted leaf tag>

Step 3. update information scent estimate in WM

Step 4. decide: if there is another highlighted leaf-tag then Locate <highlighted tag > & goto step 2

Return with goal accomplished

Judgement call: At first GridVis highlighted all tags with the same colour (whether they were section tags or leaf tags). Hence this method looked for <highlighted tag> (which requires a fast pre-attentive search) but since it is only the leaf tags which convey the information scent, a lot of time (max 50% of total) was being wasted on parent tags. The interface was therefore changed to de-emphasise the section tags by highlighting with a light grey. Hence a pre-attentive visual search could be used to locate all leaf tags.

Judgement call: The 'visual search for <highlighted leaf tags>' operator roughly locates all the highlighted tags, so that they can be subsequently located from memory without needing a full individual search for each one. An alternative approach, would be to do a new visual search for each tag. This was tried, and produces performance times that were much higher than those observed. Also pre-attentive visual searches

produce a 'pop-out' effect where all the stimuli with the feature of interest are located in parallel.

Method for goal: 3 select cell with highest profitability

Step 1. choose cell with highest profitability

Step 2. point at <cell>, verify selection, select <cell> and verify selection

Return with goal accomplished

3.4 Co-occurrence and restricted applicability

Here the applicability of the cell being considered for selection is taken into account, but that of the other tags relating to the paragraph are not.

Method for goal: 1 select the unselected paragraph cell on <tag> with the highest profitability

Step 1. accomplish goal 2 (step 1_cost = (goal2_cost*(n-k))/n)

Step 2. accomplish goal 3

Return with goal accomplished

Judgement call: It is likely each time a section is revisited, a little more of the information scent data will be held in memory. It is not clear how this discounting of cost will occur. Therefore we assume the simplest case, where the cost of executing goal 2.1 decreases linearly reaching 0s at the last tag. So goal2.1_cost(n-k))/n were n, is the total number of tags, and k is the number of un-investigated tags. Cost is only discounted for this step and goal, since the other steps could not be speeded by information in memory accrued on last visit to the section.*

Method for goal: 2 gather information scent from paragraph cell and their tags

Step 1. visual search for next unselected <cell> on <tag> row

Step 2. point at <cell> and verify get-from-task, <cell-applicability>

Step 3. store information scent of <cell> in WM

Step 4. accomplish goal 4

Step 5. Decide: If visual search for next unselected cell on <tag> row finds a cell then, goto step 2

Return with goal accomplished

Method for goal: 4 evaluate information scent of co-occurring tags

Step 5. visual search for <highlighted leaf tags>

Step 6. get-from-task <highlighted leaf tag>

Step 7. update information scent estimate in WM

Step 8. decide: if there is another highlighted leaf-tag then Locate <highlighted tag > & goto step 2

Return with goal accomplished

Judgement call: At first GridVis highlighted all tags with the same colour (whether they were section tags or leaf metadata tags). Hence this method looked for <highlighted tag> (which requires a fast pre-attentive search) but since it is only the leaf tags which convey the information scent, alot of time (max 50% of total) was being wasted on parent tags. The interface was therefore changed to de-emphasise the section tags by highlighting with a light grey. Hence a pre-attentive visual search could be used to locate all leaf tags.

Judgement call: The 'visual search for <highlighted leaf tags>' operator roughly locates all the highlighted tags, so that they can be subsequently located from memory without needing a full individual search for each one. An alternative approach, would be to do a new visual search for each tag. This was tried, and produces performance times that were much higher than those observed. Also pre-attentive visual searches produce a 'pop-out' effect where all the stimuli with the feature of interest are located in parallel.

Method for goal: 3 select cell with highest profitability

Step 1. choose cell with highest profitability

Step 2. point at <cell>, verify selection, select <cell> and verify selection

Return with goal accomplished

4 Operator algorithm for get-from-task <cell-applicability>

This derivation of the get-from-task <cell-applicability> operator is based on the method and operators from (John & Newell 1990)

P = perceptual operation (100ms), C = cognitive operation (50ms)

Operator invoked by algorithm step	Algorithm steps for doing get-from-task <cell-applicability>
P	Stimulus <- get stimulus
C	IF-SUCCEEDED is-high-applicability?(stimulus)
C	THEN applicability <- get- high-applicability-label
C	ELSE IF-SUCCEEDED Is-medium-applicability?(stimulus)
C	THEN applicability <- get- medium-applicability-label
C	ELSE applicability <- get- low-applicability-label

Table 30 Algorithm for doing get-from-task <cell-applicability>

The algorithm in Table 30 can be instantiated for each of the three states of applicability, in order to count up the number of operators invoked. The results of running the algorithm if:

app=high → P+ 2C,

app=med → P+3C,

app=low → P+3C.

Therefore the mean cost of the get-from-task <cell-applicability> operator is:

3/3P+8/3C

= P + 2.7C

= 100 + 2.7*50

= 235ms

Appendix 5 - Paragraph time coding procedure

The time spent on individual paragraphs was measured using video footage of the reader's pen movement and page turns. The time taken to accept or reject a paragraph was measured by noting down the time, on the video's time code, at which the reading each paragraph started and stopped. Start and stop times were deduced by observing the movement of the reader's pen and page turns. A start time is taken at the point where the pen's nib enters or comes horizontally into line with a paragraph. Correspondingly, a stop time is taken when the pen's nib leaves or comes horizontally out of line with the paragraph. Interpreting the pen movement is in fact more subtle then this; it is covered in more detail below. When the reader begins to turn a page, which ends at a paragraph border, this is taken as a stop time. A start time is then taken as soon as the next page is in view. However if the page ends before the paragraph has been completed, the start and end of the page turn are noted and used to subtract out the page turning time from the paragraph reading time.

The procedure for measuring paragraph reading times, summarised above, is described in more detail below via a number definitions and procedures particular circumstances.

Time spent on a document – this starts once the table of contents is in view, and stops once the participant has turned over the last page of the document and the text is no longer in view.

Start time for a paragraph –

- The start time for a paragraph begins when the pen's tip is vertically aligned with the top of the paragraph, or, if the paragraph is the first in an article, the title.
- If the paragraph is at the start of a new page, assign a start time once the pen has visibly started its trajectory towards it.
- The time taken to read titles is included in the time to read the first paragraph since it is often not possible to distinguish the reading of the title from that of the paragraph using the pen position.
- When there is a ballistic/parabolic swing of the pen between paragraphs the junction between the time allocation is in the middle of the swing or at the high point. This is because although the swing would seem to indicate a saccade to the next paragraph/heading, it might be representing a skim over the paragraph.

Stop time for a paragraph –

- A stop time is taken when the pen's nib leaves or comes horizontally out of line with the paragraph
- If the pen has not been used for a paragraph, indication of an end can be inferred from movements/gestures made with the pen. For example if upon a page turn the pen is placed near the top of the page and then after a short while it is moved up a little, and a short while later the reader starts to highlight the following paragraph. The 'moved up a little' gesture can be taken as a sign of the reader moving on to the following paragraph.

Page turns –

- If there is a page turn in the middle of a paragraph the time taken for the page turn is subtracted from the total paragraph time (since the cost estimates for

paragraphs reading is for onscreen reading). This only needs to be done when the reader evidently spends time on the portion of the paragraph that lies over the page.

- Page turns are measured from the beginning of page lifting to the point where the next page is fully visible to the video camera. If the reader's hand comes down to turn the page, such that it must be obscuring their view of the text, then the page turn starts as soon as the text is obscured (from the reader's point of view).
- If there are multiple page turns for one paragraph, these are inputted, summed, and a notional single page turn time is created to give the total page turning time.
- When there is no clear delimitation between the one or more paragraphs, the time spent is divided equally among them.

Skipped paragraphs –

- If the page is turned quickly without settling down to any of its paragraphs, then count these as missed paragraphs.
- It is often unclear whether an article/paragraph has been skimmed or missed. If a paragraph is part of an article and the previous paragraph has obviously been looked at (i.e. the pen rests by it), but the pen is then moved parabolically (i.e. up and over) the next paragraph(s). Then it can be safely concluded that this paragraph has been skipped.

1 Codes assigned to paragraphs

As well as noting the time spent on each paragraph, the transcription process involved assigning several codes to each paragraph. Primarily these indicated whether these

paragraphs were highlighted, and thus deemed relevant, or not. An additional code was included in the hope of excluding the least reliable paragraph times in order to reduce the variability of the paragraph time estimates. Essentially, paragraphs were coded as 'unclear' when there was reason to believe that the time measurement was not accurate or representative. This occurred, for example, when start and stop times were not clear from the pen movements, or when the subject was distracted from the task. Paragraphs were given page a 'border' code, if they were at the very top or bottom of the page. This was done because there was a greater degree of uncertainty about paragraph times at page borders since it is not clear how soon reading begins after the page comes into view.

More specifically the 'not clear' code was applied when:

- The time spent on one or more paragraphs cannot be distinguished and they are of different sizes.
- It also applies when it is not clear whether paragraphs have been skipped or read.
- When the experimental procedure has not been followed – e.g. one participant marked some of the paragraphs with question marks.

Appendix 6 – Graphs showing full data for the effect of document length in the paper performance experiment

This Appendix shows graphs of data from the paper performance experiment. In particular they show the effect of document length, displaying individual data points as well as mean task times. These allow the reader to better appreciate the extent of overlap between scores with documents of differing length. They were not included in the main body of the thesis, due to the large amount of space they require.

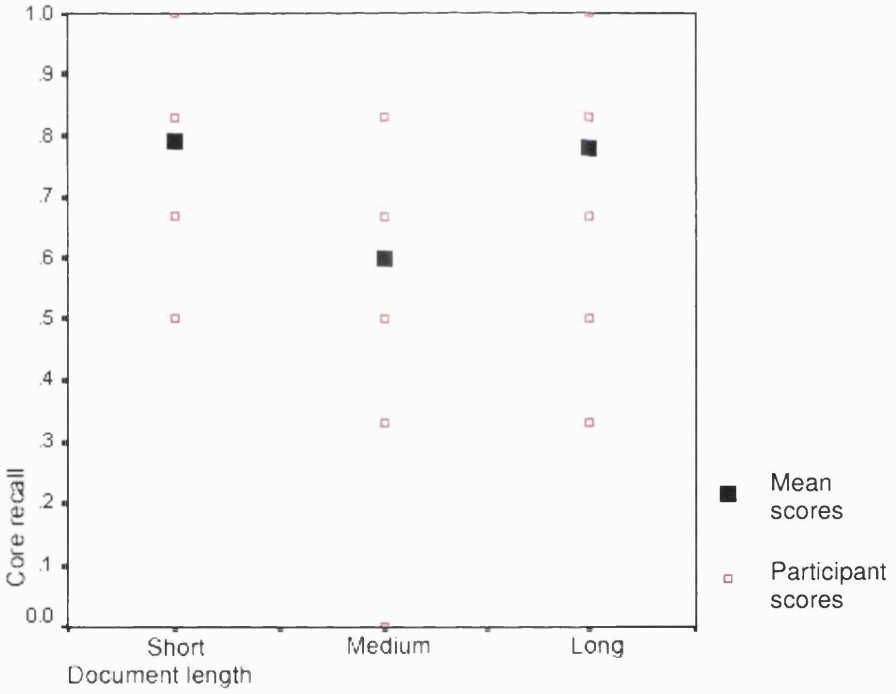


Figure 92 The individual participant times and the mean scores, for core recall against document length

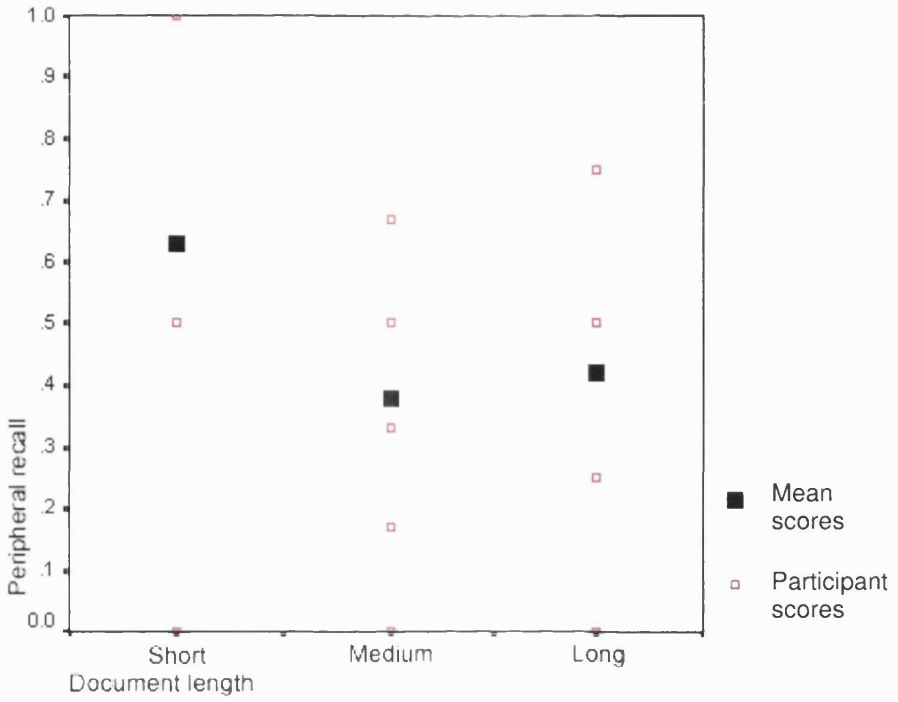


Figure 93 The individual participant times and the mean scores, for peripheral recall against document length

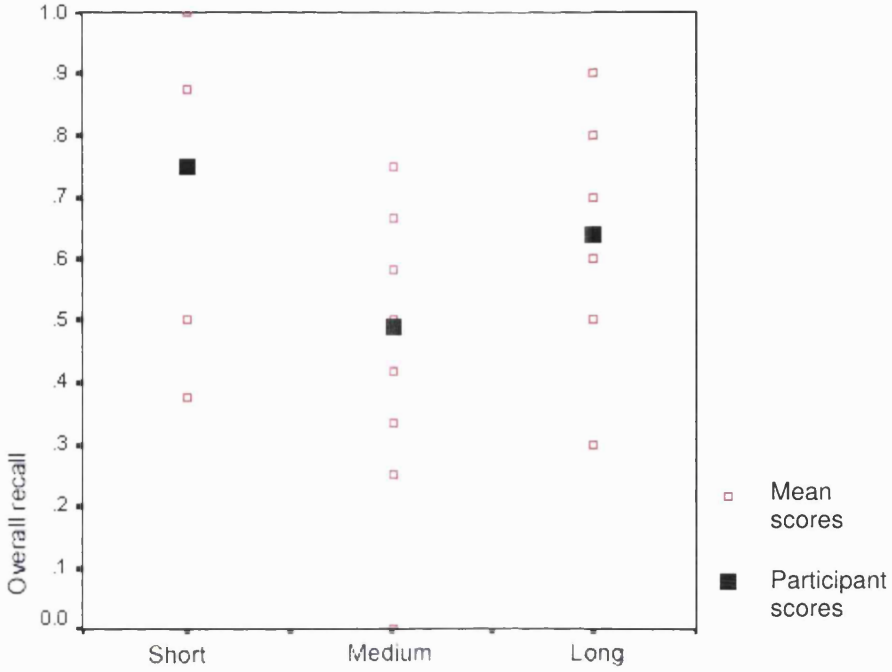


Figure 94 The individual participant times and the mean scores, for overall recall against document length

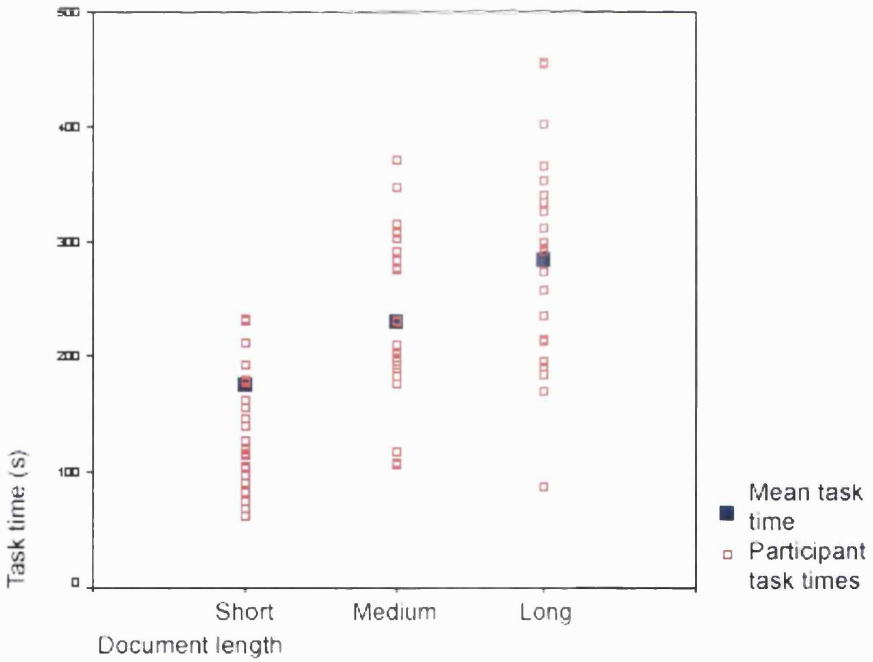


Figure 95 The individual participant times and the mean scores, for task time against document length

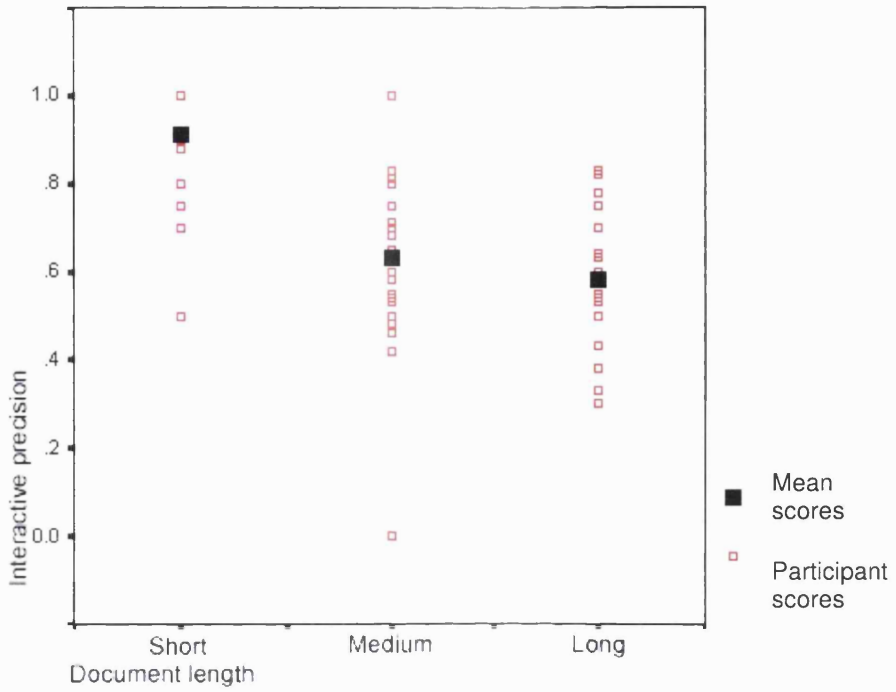


Figure 96 The individual participant times and the mean scores, for interactive precision against document length

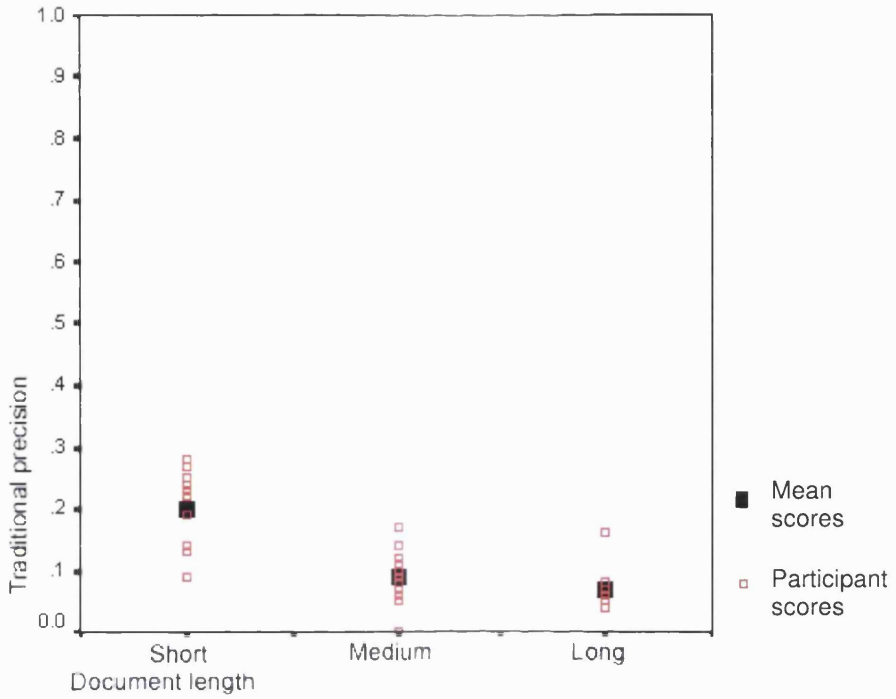


Figure 97 The individual participant times and the mean scores, for traditional precision against document length

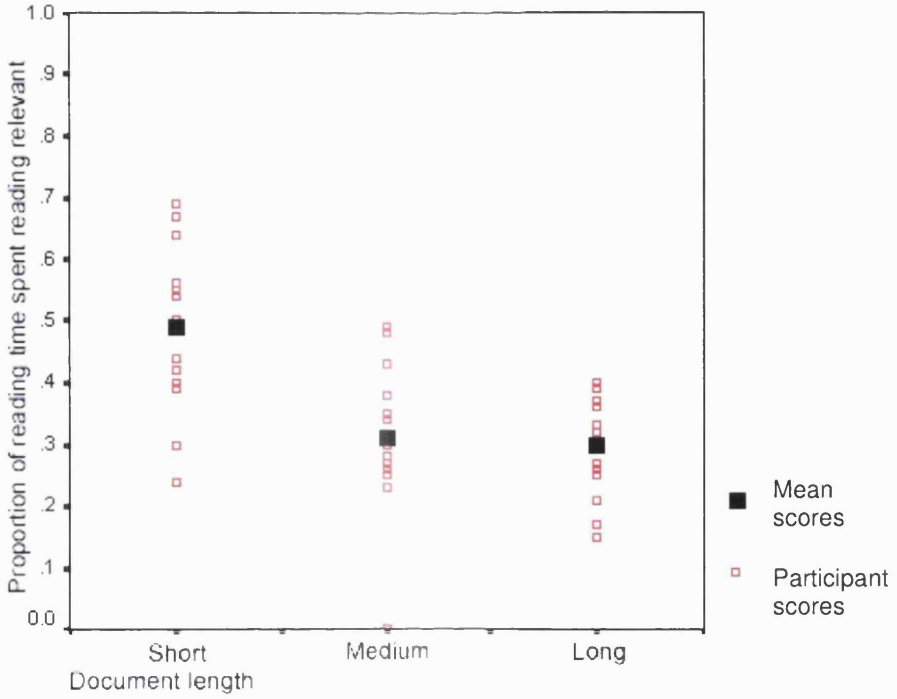


Figure 98 The individual participant times and the mean scores, for proportion of reading time spent with relevant against document length

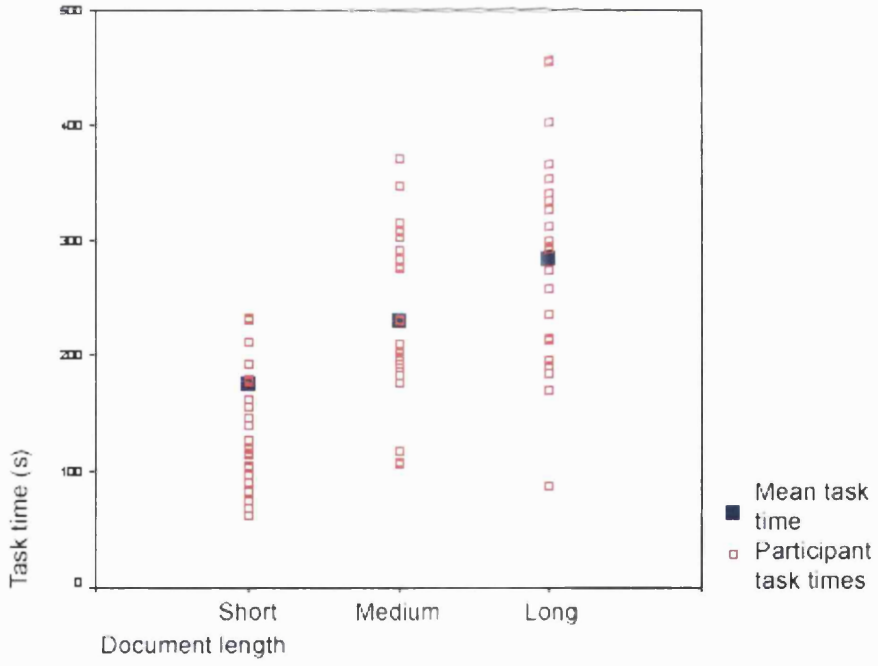


Figure 99 The individual participant times and the mean scores, for task time against document length

Appendix 7 Further details on the documents used in the best-case analysis

1 Setting and measuring document length

It was felt that document size would have a negative effect on performance with paper, but not with GridVis. Thus the three documents used were varied in size, ensuring that an even spread of sizes was achieved across a number of different measures. The sizes of documents used, according to each of the measures used, are shown in Table 31 below.

The first three measures of document lengths are fairly common, the fourth needs some further explanation. It arises from the fact that the probability of stumbling upon relevant material is an important factor behind the effect of document length on paper document search difficulty. As document length was increased, the proportion of relevant paragraphs was decreased and with it the probability of randomly happening upon a relevant paragraph. Therefore search becomes harder as documents get longer. Hence without paying careful attention to this factor, it would be possible to vary the length as desired without achieving the same variation in the prevalence of relevant paragraphs and hence search difficulty. The documents were therefore constructed such that the proportion of relevant paragraphs was varied appropriately.

Document	No of pages	No of paragraphs	No of words	Proportion of relevant paragraphs
Long	14	60	7161	0.083
Medium	8	39	4059	0.154
Short	5	16	2181	0.25

Table 31 Various measures of different the lengths of each of the three documents used

Clearly, 'the proportion of relevant paragraphs' is only a crude indicator of the probability of randomly locating a relevant paragraph. Other factors could be considered such the proportion of relevant words, text height, and sub-sections. Factors such as the position on page and distance from the nearest heading may also effect how easy it is to locate relevant information in a document. However, paragraph number was singled out for attention since this can be controlled simply and would seem to be most clearly related to the effect of document length.

2 Other considerations in the document construction

The documents were constructed to hold constant all factors other than those intentionally being varied.

In order to promote awareness about the document organisation, contents pages were added. The industry update documents are divided up into a number of sections such as 'e-commerce' and 'supply chain'. These sections could be very useful in locating relevant information. Some participants in the paper performance experiment might be more aware of these sections and thus do better than others. It was decided to deal with this problem by including the sections and their locations on a contents page attached to the front of each document.

In order to minimise the effect of answer location on performance across conditions, the locations of relevant paragraphs was controlled. To take an extreme case, if answers were always in the document's first articles, performance would be faster than otherwise. A simple procedure for placing relevant paragraphs was defined so as to minimise this problem as far as the constraints of document type (i.e. the location of sections) would allow. Relevant articles were evenly spread within a section and

surrounded by non-relevant articles. When there was only one relevant article in a section it was located in the middle of the section.

Finally in order to create documents representative of the document type, the sections were constructed with the same order and relative proportions as the original 'industry updates'.