An Intelligent Hypertext System

Michael Junke Hu

Thesis submitted for the degree of
Doctor of Philosophy
of the University of London.
November 1994

Department of Computer Science
University College London
ABSTRACT

This thesis investigates the applications of machine intelligence in information generation, organization, manipulation, search and retrieval.

In order to alleviate and solve some problems in the present information retrieval (IR) community, and to increase the efficiency and effectiveness of information systems, a new data structure is proposed in the thesis. The conceptual index is external to the collection of information components (documents). It integrates the conventional global index with a special semantic network. As a result, a much richer set of concepts, as well as the relationships between concepts, can be represented in the data structure. It is shown in the thesis that such a data structure is more suitable for sophisticated IR environments such as hypertext, and could make automatic information generation, self-adjustment and evolvement, inferencing and reasoning possible.

Based on the conceptual index, a soft-link hypertext model is developed and investigated. The soft-link hypertext model covers the Boolean search model, the probability model, the traditional hard-link hypertext model and the soft-link hypertext in one infrastructure. Its main features include automatic generation of the conceptual index, self-adjustment and evolvement, user-centered services for information retrieval and applications of machine intelligence in all aspects of the model.

The soft-link hypertext model, including its state-space and all operations occurring in the space, is fully presented and evaluated in the thesis. It is implemented in a soft-link hypertext system called the Enhanced SuperBook and assessed in several small-scale controlled IR experiments. Its strengths, weakness, similarity with, and difference from other information models and systems are also studied in depth.

Based on such an investigation, the thesis concludes that many aspects of information retrieval should benefit from the extensive application of the machine intelligence. Automation in information generation, organization, self-adjustment and evolvement, and assistance for information retrieval can not only increase the efficiency and effectiveness of the information systems, but also represents the essence of, and should have great impact on, future generations of IR systems.
ACKNOWLEDGEMENTS

During the three-year Ph. D period I have constantly received support, both professionally and emotionally, from various people. Without them, the work presented in this dissertation would not have been possible.

Firstly, I am very grateful to my supervisor, Professor Peter T. Kirstein, for the invaluable encouragement and guidance I received from him during the last 36 months. Since my first day in UCL, I have benefited enormously from his supervision of my research and the presentation of technical results. He has provided great stimulation and inspiration. I am indebted to Professor Kirstein for giving me so much of his time and knowledge, which has encouraged me greatly.

I would also like to express my deepest appreciation to my supervisor, Mr. Nigel Chapman. I like and respect Nigel both as a supervisor and as a friend. It is he who has given me much important professional advice and pushed me to make progress towards my final goal.

My implementation of the Enhanced SuperBook was based on software developed by Professor Michael Lesk and his team in Bellcore. I am indebted to him for his patient support for technical details.

My sincere thanks go also to all my friends and colleagues in the Computer Science Department at UCL for their friendship and advice. In particular, I am very grateful to Mark Levene, Goli Montasser-Kohsari, Russel Winder, Paul Barker, Jon Crowcroft and Angela Sasse, to Jonathan Crompton and Steven Davey in the System Group, and to Annlette Riley, Kevin Randall, Dawn Bailey and Natalie May.

My Ph. D research was sponsored by the British Council and the SBFSS. I also want to take this opportunity to express my sincere gratitude to the British Council, and to my programme officer, Ms. Grace Persuad, in the British Council.

A lot of friends helped me with my experiments, both as domain experts and information users. I want to thank them all for their time and effort.

Finally, and most importantly, special thanks to my parents and my brother for their encouragement and support throughout the last three years, and to my close friends, Anthony Lavender and Mark Story, for their love, companionship and kind support.
CONTENTS

ABSTRACT 2

ACKNOWLEDGEMENTS 3

1. Introduction 13
   1.1 Background 13
   1.2 My Research Work 14
   1.3 The State of Art 15
   1.4 Notations Adopted 15
   1.5 Outline of the Dissertation 16

2. Information and Information Retrieval 18
   2.1 Information Space 18
   2.2 Information Retrieval 19
      2.2.1 Traditional IR 21
      2.2.2 Hypertext 25

3. Hypertext 27
   3.1 Definition 27
   3.2 History 27
      3.2.1 The First Generation Hypertext 28
      3.2.2 The Second Generation of Hypertext 29
      3.2.3 The Third Generation of Hypertext 30
   3.3 Why Hypertext? 32
   3.4 Models of Hypertext 34
      3.4.1 Models of Hypertext 34
      3.4.2 Problems Existing in Present Hypertext Systems 35
   3.5 Soft-Link Hypertext 39
6.6.3 The Unique Characteristic and Its Generalization .................................. 89
6.7 Summary ............................................................................................................... 90

7. Intelligent Mechanisms in the Soft-Link Hypertext Model ......................... 91
  7.1 Uncertainty ......................................................................................................... 91
  7.2 Uncertainty in IR .............................................................................................. 92
    7.2.1 Traditional Search and Retrieval ............................................................. 92
    7.2.2 Traditional Hard-Link Hypertext .......................................................... 93
    7.2.3 Soft-link Hypertext ................................................................................. 94
  7.3 User Modelling in Soft-link Hypertext ............................................................. 95
    7.3.1 User Modelling ..................................................................................... 95
    7.3.2 User Modelling in Soft-link Hypertext ............................................. 97
  7.4 Learning in Soft-link Hypertext ...................................................................... 99
    7.4.1 Long-term Memory ............................................................................. 100
    7.4.2 Short-term Memory ........................................................................... 100
    7.4.3 Relation between STM and LTM .................................................... 101
  7.5 Reasoning in Soft-Link Hypertext ................................................................. 102
  7.6 Intelligence in the Soft-link Hypertext Model ................................................. 105
  7.7 Summary .......................................................................................................... 106

8. An Intelligent Information System ................................................................. 107
  8.1 Overview of the Model ................................................................................... 107
  8.2 The Soft-link Hypertext Model — Description and Novel Attributes ....... 109
    8.2.1 The Dexter Hypertext Reference Model ............................................. 109
    8.2.2 The Soft-link Hypertext Model ......................................................... 111
  8.3 Features of the Model .................................................................................... 114
    8.3.1 Summary .............................................................................................. 117
  8.4 Evaluation of the Model ............................................................................... 118
    8.4.1 Generation of Data Structure ............................................................ 118
    8.4.2 Applications of the Model in IR ....................................................... 118
8.4.3 Information Introduction ........................................... 119
8.4.4 Information Expansion ............................................. 120
8.4.5 Information Pinpointing ........................................... 120
8.4.6 Summary .............................................................. 121
8.5 Implementation of the Soft-link Hypertext Model .................. 121
  8.5.1 SuperBook .......................................................... 121
  8.5.2 Implementation of the Model ................................. 125
  8.5.3 Databases ......................................................... 127
  8.5.4 The Browser ...................................................... 128
  8.5.5 Interfaces .......................................................... 128
8.6 Conclusion ............................................................... 134

9. Evaluation of the System ................................................. 135
  9.1 Evaluation ............................................................... 135
  9.2 Evaluation of the Model ............................................. 136
  9.3 Evaluation of the System ........................................... 136
    9.3.1 The Variables Measured ..................................... 138
    9.3.2 Methodology .................................................... 139
    9.3.3 Design of the Experiments ................................. 139
  9.4 Experiment 1 — Index Formulation ............................... 142
    9.4.1 Purpose and Design .......................................... 142
    9.4.2 Experimental Process ....................................... 143
    9.4.3 Experimental Results ........................................ 144
    9.4.4 Analysis of Results .......................................... 146
    9.4.5 Conclusion ...................................................... 148
  9.5 Experiment 2 — Soft-link Hypertext Browser .................... 149
    9.5.1 Purpose and Design .......................................... 149
    9.5.2 Experimental Process ....................................... 150
    9.5.3 Experimental Results ........................................ 151
    9.5.4 Analysis of Results .......................................... 159
# LIST OF TABLES

8.1 Evaluations of IR models ................................................. 121
9.1 Index terms in Group 1 ............................................. 144
9.2 Index terms in Group 2 ............................................. 145
9.3 Decreasing percentage of recall and precision ................. 146
9.4 Distribution of the subject in two treatments ................. 149
9.5 Distribution of treatments and subjects in Experiment 2 .... 150
9.6 Search time (minutes) spent on each keyword (Group 1) .... 151
9.7 Search time (minutes) spent on each keyword (Group 2) .... 152
9.8 User actions needed for search and retrieval (Group 1) .... 154
9.9 User actions needed for search and retrieval (Group 2) .... 155
9.10 Recall & precision for Group 1 ................................... 157
9.11 Recall & precision for Group 2 ................................... 158
9.12 The number of pieces of information obtained by the user. 166
9.13 The accuracy of answers for Task 1 (free browsing) .......... 166
9.14 Number of User actions needed for Question 1 ............... 170
9.15 Time (minutes) spent on Question 1 ......................... 170
9.16 Number of User actions needed for Question 2 ............... 175
9.17 Time (minutes) spent on Question 2 ......................... 175
9.18 Number of User actions needed for Question 3 ............... 178
9.19 Time (minutes) spent on Question 3 ......................... 178
9.20 Number of User actions needed for Question 4 ............... 181
9.21 Time (minutes) spent on Question 4 ......................... 181
LIST OF FIGURES

1.1 The flowchart of the dissertation .............................................. 17
2.1 Information system environment .............................................. 19
2.2 Functional overview of information retrieval ............................ 20
2.3 Functional overview of hypertext ............................................. 25
3.1 The three layers of the Dexter model ........................................ 34
3.2 Redisplay of Figure 2.3 ............................................................ 36
3.3 A soft-link hypertext model .................................................... 40
3.4 The difference between three different IR models ....................... 41
3.5 The conceptual index in the soft-link hypertext model .................. 42
4.1 A conceptual graph ................................................................. 45
4.2 A conceptual index in the soft-link hypertext model ..................... 47
4.3 The correspondence of different layers ...................................... 48
5.1 Resolving power of significant (medium-frequency) words .............. 51
5.2 Local and global approaches .................................................. 54
5.3 A conceptual index — global + local approaches ......................... 55
5.4 A semantic net supports the local associations ......................... 56
5.5 Each concept has an attribute attached to it ............................... 56
5.6 Resolving power of significant words for the semantic net .......... 59
6.1 The typical three-layer connectionist network in IR ..................... 65
6.2 A connectionist network defined on the Index Space .................. 68
6.3 The conceptual index represented as a connectionist network ....... 70
6.4 The connectionist network in the Index Space ........................... 72
6.5 A Kohonen layer ................................................................. 85
6.6 The Kohonen layer in the conceptual index ............................... 87
7.1 Intelligent mechanisms in the soft-link hypertext model ............... 105
8.1 The soft-link hypertext model vs. the Dexter model .................... 107
8.2 The structure of the Index Space ............................................. 108
Chapter 1
Introduction

This chapter presents a framework of my research work, including the problems addressed, the hypotheses examined and the notations adopted. It also gives a brief description of following chapters in the dissertation.

1.1 Background

A massive, global network of information through which people can browse freely with the aid of electronic technology has been envisaged for decades. Ever since Bush proposed the idea in 1945 [27], hypertext has increasingly become a very important tool for information organization and information retrieval. Many hypertext systems have been in operation since the 1980s and the idea behind hypertext has been enticing generations of researchers to continue to explore the field. However, this miracle of global information network is still around the corner. Some of the following problems, which hinder the further development of hypertext, have been discussed widely, but have yet to be solved.

1. What is the proper configuration of hypertext? What kind of data structure should be used to represent this network of information?

2. How can this network of information, with a large number of intra- and inter-information links and associations, be generated? What kinds of tools can be used for the generation process?

3. What is the proper relationship between searching by index and searching by associations? What effects do they have on the information retrieval and information users?

4. What role does the machine intelligence play in information retrieval? How can it help the human intelligence, the human thinking associations and information retrieval?

5. How should an intelligent information system be evaluated?

These are the problems to be addressed in this dissertation.
1.2 My Research Work

The hypothesis that I examine in the dissertation is — A conceptual index, which constitutes a semantic net upon individual index terms of the traditional global index, supports both selection by index and selection by associations. Compared with traditional approaches, the conceptual index presents an improved data structure for information retrieval. An information retrieval system based on such a structure (i.e. a run-time soft-link hypertext system studied here) can offer a graceful model and infrastructure for information organization and presentation, and improve features of the existing information retrieval models and systems. Thus, it provides effective solutions to the problems listed in Section 1.1.

Furthermore, the hypothesis can be divided into several sub-hypotheses:

1. The conceptual index could be built up automatically, or semi-automatically;
2. The conceptual index could present a more consistent and systematic approach to organization of the information network;
3. Whilst the index structure could facilitate global access to any index terms (concepts), the semantic net supports the relationship amongst different index terms (concepts), and therefore, constitutes local associations among different pieces of information. Information systems based on such a structure could be user-centered and become more efficient and effective;
4. Machine intelligence could be explored to generate such information networks, improve the quality of its organizations and presentations, and finally, help users to effectively retrieve information. Such an intelligence may include the following three different, but related elements:
   - A better user model could be built up by an intelligent system;
   - Machine memory (learning) could be used to record the past experiences, and to self-adjust and improve the network based on the experiences;
   - A reasoning mechanism could be used to reason and make inferences on the user’s information needs.

The work presented in following chapters of this dissertation sets out to prove the above hypothesis. A model is developed to show its validity on a theoretical level, whilst an implemented system demonstrates it and provides a practical evaluation.
1.3 The State of Art

The research presented in this dissertation is based on the following concepts:

1. The theoretical studies and practices in information retrieval (IR), which have been very well summarized in [37] [134] and are reviewed in Chapters 2 and 3;

2. Advances in artificial intelligence (AI), especially those concerning the application of AI in information retrieval. The related branches include Expert Systems [114], knowledge-based systems [45] and the connectionist models [54];

3. Finally, the model and implementation of an integrated soft-link hypertext system are based on SuperBook, which is a full-text information retrieval system from Bellcore [56] [57]. The database used is the electronic version of journal, Environmental Science and Technology (1991), from the American Chemical Society. A more detailed description of SuperBook, including its functions, its browser and interfaces, is given in Section 8.5.1.

1.4 Notations Adopted

In order to formally and concisely present the soft-link hypertext model developed in the dissertation, the following notations are used.

1. Conceptual graph and conceptual structure (Chapter 4):

   The conceptual index introduced in the dissertation is represented as a directed graph called the conceptual graph. With its node representing the concepts and arcs the connections (links) between concepts, the conceptual graph can help in presenting the state-space of the soft-link hypertext model. Furthermore, once the conceptual graph is introduced, all notions in the graph theory can be easily adopted to explain the dynamics of the model.

2. Formal specification of Z (Chapter 6 and 7):

   The detailed description of all events happening in the soft-link hypertext model is represented as a formal specification in the Z notation. These events include the changes of the state-space and its applications in information retrieval. Rules used to control these events are presented first as separate schemas, which are combined together to constitute the whole model.
1.5 Outline of the Dissertation

This dissertation consists of five main parts (Figure 1.1): analysis of problems (Chapter 2 and 3), development of the soft-link hypertext model (Chapter 4, 5, 6 and 7), its implementation (Chapter 8), experiments and evaluations (Chapter 9) and conclusion (Chapter 10). The contents of the main chapters are set out as follows:

1. Chapter 2 surveys the models and systems currently used in information retrieval. These include the Boolean search model, the vector space model, the enhanced Boolean search model and the probabilistic model;

2. Chapter 3 provides a more detailed survey on hypertext: its definition, its history, the models used in hypertext and their characteristics, etc. The idea of soft-link hypertext is introduced at the end of the chapter;

3. The conceptual structure is defined in Chapter 4. It provides a formal notation and theoretical basis for further presentations in Chapters 6, 7 and 8;

4. Chapter 5 deals with the conceptual index and its formulation. The traditional indexing methodology is first reviewed and used to formulate the mapping from individual index term to information components. Then, a mechanism is introduced to generate the special semantic net needed for the conceptual index;

5. Chapter 6 presents the connectionist network adopted in the soft-link hypertext model, and its activation and learning rules. The connectionist network is used for controlling the dynamic operations in the soft-link hypertext model;

6. Based on the connectionist network, several intelligent mechanisms are developed in Chapter 7. These different mechanisms are combined together to constitute the intelligent soft-link hypertext model;

7. The soft-link hypertext model, as well as an implemented system (called the Enhanced Superbook), is illustrated in Chapter 8. It is evaluated by comparing with other information-retrieval models and systems;

8. The experiments in which the hypotheses are examined in several small-scale controlled IR environments are reported in Chapter 9. Four experiments are designed in the research, to test and assess different features of the system;

9. Finally, the conclusions are reached in Chapter 10. The proposals for further work, as well as the impact of the soft-link hypertext model on future research of global information systems, are also discussed in the chapter.
17

Introduction
1. Problems; 2. Hypothesis; 3. Layout of dissertation

Models of IR
Models of hypertext & unsolved problems
A soft-link hypertext model

Conceptual structure
Automatic hyperization
User modelling
Learning
Reasoning

The soft-link hypertext model
Evaluations & implementation

Experiment 1
Experiment 2
Experiment 3
Experiment 4

Evaluation of the implementation based on experimental results

Conclusion

Chapter 1
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Chapter 6 & 7
Chapter 8
Chapter 9
Chapter 10

Figure 1.1: The flowchart of the dissertation.
Chapter 2

Information and Information Retrieval

This chapter presents a general view of the information space and information retrieval systems. Different models currently used in information retrieval (IR) are reviewed here. They include the Boolean search model, the vector space model, the enhanced Boolean search model, the probability model and hypertext. The conclusion of the chapter is that hypertext provides a new and higher-level model to organize and retrieve information.

2.1 Information Space

Much of the information circulating in the modern world consists of written text, figures, tables, chemical structure formulae, sound, picture, etc. Often the information is available in machine-readable form; it can be stored and reproduced automatically; finally, the information can be transmitted from place to place via electronic networks.

It is easy to store large masses of information. However, storage in itself is of no value unless systems are designed which make selected items available to interested users. The following operations are needed, in order to make this selection process efficient:

1. The information components must be analyzed; appropriate content identifiers must be generated and attached to the stored components;

2. The user needs must be identified; they must be formulated in forms understandable by an automatic system;

3. Representations of information components and user needs must be compared, leading to the retrieval of information components judged to be sufficiently close to the information requested.

Such an automatic system, including all the information and auxiliary tools, is often referred to as an Information Space. The process in which some useful information is extracted from an Information Space is called information retrieval.
2.2 Information Retrieval

Information retrieval (IR) is concerned with the representation, storage, organization and access of information. Usually, an information retrieval system consists of a set of information components, or documents (DOCs), a set of queries made by the information user (QUEs), and some mechanism (SIMILAR) which is used to determine which, if any, of the information components meets the requirements of the queries (Figure 2.1) [134]. The three main elements of IR are as follows:

![Figure 2.1: Information system environment.](image)

1. Representation of the Information Space;

An Information Space refers to all the information stored in an information retrieval system (Figure 2.2). It usually includes the following elements:

- All the information components stored (the Document Space);
- All auxiliary knowledge (e.g. the structure of those documents, the domain, its users and their problems, etc.) and tools needed for storing, organizing and accessing these information components (the Index Space).

2. Representation of the information user;

Information users are people who access information systems in order to find some needed information (Figure 2.2). They represent diverse backgrounds and information interests, with different levels of interacting skills and different abilities to specify their information needs either explicitly or implicitly.

The user's information needs are often represented by a set of queries (QUEs).
3. Information search — comparison of above two representations.

Some kind of SIMILAR mechanism is used to compare these two representations and select, among the information components in the Document Space, the ones which meet the requirements of the user's information needs.

In theory, the selection between queries and documents is obtainable by direct comparison, as suggested in Figure 2.1. However, this direct comparison is not always practical. It becomes far too slow when the Information Space is large. Instead, a data structure is usually introduced in the Index Space, which includes a set of content identifiers and provides a more systematic view of the Document Space; such a data structure can help in determining the similarity between those two representations (Figure 2.2). The idea is that, once the proper content identifiers can be found to match the information needs, the related documents can be retrieved by following the pointers provided by the data structure.

![Diagram of Information Retrieval](image)

**Figure 2.2: Functional overview of information retrieval.**

Traditionally, the mapping from the Document Space to the Index Space is called **indexing**, that from the queries to the Index Space **search formulation**, and the comparison process **searching**.

Several models have been used in indexing, search formulation and searching.
2.2.1 Traditional IR

1. Characteristics of traditional IR;

In traditional information retrieval, the data structure is usually based on a set of keywords or phrases called index terms. Index terms may include a controlled list of words or all words in a Document Space (full-text index). All information components in the Document Space are identified by these index terms. The indexing process is to choose, either manually or automatically, the set of index terms which could represent most effectively the characteristics of each individual document and build the proper associations (mappings) from these index terms to the related documents. Users' queries are also expressed by these index terms. Therefore, the judgement of relevancy is based on the index terms shared by the queries and those relevant information components.

2. Depending on the searching process (SIMILAR mechanism) used in the models, traditional IR can be divided into four groups: the Boolean search model, the vector space model, the enhanced Boolean search model and the probability model. These models can be characterized as follows:

- **Boolean Search Model**;

  The Boolean search model is a simple but efficient model on which most commercial IR systems are based. Its comparison process operates on a one-to-one basis between a query term and the index term. Index terms are usually stored alphabetically in files called inverted files. Boolean operators are applied to formulate more complicated queries. The advantages of the Boolean search model are listed as follows:

  (a) It is a good fit for a very simple and clear mathematical model. All technologies related to the model are well understood;
  (b) Its implementation is relatively easy. Running a Boolean search system can be straightforward and very efficient;
  (c) Very simple term relationships (like synonymous) and simple phrases are expressible with Boolean operators;
  (d) All processes, including indexing, search formulation and searching, can be automated;
  (e) With some added features like term weight and user feedback, a high degree of effectiveness is sometimes obtainable.
The main problem of the Boolean search model is the limitation of expression of the Boolean logic. Among the disadvantages are:

(a) The use of Boolean operators may prove difficult for the user who is not familiar with Boolean logic;

(b) Conventional Boolean logic treats all terms as equally important and, as a result, considers all retrieved documents as equally useful;

(c) Boolean operators (AND, OR, NOT) is unusually rigid in a retrieval setting;

(d) The SIMILAR mechanism used in the model is a crisp mechanism, i.e., a document is identified as either in a retrieved set, or not, but not somewhere between.

- Vector Space Model;

Instead of assuming that all terms are equally valuable, the vector space model uses term weights to assign importance indications to index terms. If \( t \) distinct terms are available for content identification, a document \( D_i \) is representable internally as a \( t \)-dimensional vector of pairs, \( D_i = (d_{i1}, w_i^{d_1}; d_{i2}, w_i^{d_2}; \ldots; d_{it}, w_i^{d_t}) \), where \( d_{ij} \) represents the \( j \)th term assigned to document \( D_i \) and \( w_i^{d_j} \) is the corresponding term weight. In principle, all \( t \) terms could appear in each vector: a weight of zero would be used for terms not present. Larger weights, between 0 and 1, would designate terms actually assigned to the document.

In the vector space model, the Boolean queries are replaced by weighted term sets similar to those used for the document representations. Thus a query \( Q \) appears as \( Q = (q_1, w_i^{q_1}; q_2, w_i^{q_2}; \ldots; q_t, w_i^{q_t}) \), where once again a weight of zero is used for terms that are absent. When both the stored documents and the information queries are represented by weighted term vectors, a global and composite vector comparison can measure the degree of similarity between a query-document pair on the basis of the weights of the corresponding matching terms. The cosine measure of similarity, computed as the normal inner product between vector elements normalized for vector length is shown in Equation (1); it has been widely used for this purpose [138].

\[
sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_i^{q_j} w_i^{d_j}}{\sqrt{\sum_{j=1}^{t} (w_i^{q_j})^2 \sum_{j=1}^{t} (w_i^{d_j})^2}}
\] (1)
The advantages of the vector space model are as follows:
(a) It is based on a clear and well-understood mathematical model;
(b) It offers simple and parallel treatments for both queries and documents;
(c) It accommodates weighted terms and therefore provides ranked retrieval output in decreasing order of query-document similarity;
(d) Query and document vectors could be easily modified, as required for query reformulation and other purposes.

Its disadvantages include the following points:
(a) As the vector space is assumed to be orthogonal and all terms are linearly independent, the model is not able to express any relationships between terms;
(b) It lacks theoretical justification for some of the vector manipulations, such as the use of the cosine measure to obtain vector similarities;

- Enhanced Boolean Search Model;

The enhanced Boolean search model accommodates term weights assigned to both query and document terms as well as strictness indicators known as p-values that are attached to Boolean operators. With this p-value, the power of expression by Boolean logic is extended. A typical query formation in the extended Boolean system would be \( [(T_1, a) \ OR^{p_1} (T_2, b)] \ AND^{p_2} (T_3, c) \), where a, b and c are the weights for terms \( T_1 \), \( T_2 \) and \( T_3 \), respectively; \( p_1 \) and \( p_2 \) are p values that control the strictness of interpretation of the Boolean operators. Values of \( p \) range from 1 to \( \infty \); the upper limit represents total strictness of interpretation, equivalent to a standard Boolean system, whereas the lower limit represents total relaxation, equivalent to a vector processing system where the distinctions between AND and OR are lost. The enhanced Boolean model covers the vector space model, Boolean search model and fuzzy-set retrieval in a common framework; it produces vastly improved retrieval performance over simple Boolean operations at the cost of a substantially increased computational effort [138].

- Probability model.

The probabilistic search model differs from those discussed previously in that it represents an attempt to set the retrieval problem on firm theoretical foundations. Here, it is necessary to estimate for each document D,
with respect to $Q_j$ the quantity $P(\text{Rel}|Q_j, D_i)$, the probability of relevance of $D_i$ with respect to $Q_j$. One approach to this estimation process regards information retrieval as an inference, or an evidential reasoning process, where an answer to a user query is deduced from the evidence provided by each document [138].

3. The advantages and disadvantages of traditional IR models.

The traditional IR models represent the oldest, best understood and most popularly used tools in IR. They are usually based on clear and solid mathematical theories. Automation is involved in all aspects of these traditional IR models; it covers indexing, search formulation and searching. The implementations of these models are comparably simple and suitable for computer applications. Finally, a reasonable good retrieval result is sometimes obtainable.

However, problems do exist with these traditional models, namely:

- They assume that user's queries are independent and ignore the psychological nature of the information user's thinking associations. As a result, they neglect the associative information enclosed in multiple queries;
- The search formulation in these models expects explicit queries from the information user. They ignore the fact that such an explicit statement of the information needs (known as recall) could be sometimes very difficult. Human beings usually feel more comfortable in another cognitive activity, called recognition. Unfortunately, recognition is not generally encouraged in traditional approaches. Semantic networks, or thesaurus, are adopted in some more recent IR systems to compensate the defective side. Nevertheless, it still remains as a big problem for the traditional IR models;
- The use of keywords (index terms) as content identifiers limits the ability of expressions in the traditional IR models:
  (a) Index terms are only able to express simple and individual concepts. They are not adequate for expressing composite concepts, or phrases;
  (b) No semantic, nor logical relationships, among different concepts, or among different information components, can be expressed;
  (c) A data structure based on such index terms are only suitable for a group of bibliographic search and retrieval, but inadequate for more complicated IR tasks. This is discussed further in Section 8.4;
- They represent a system-centered model, instead of a user-centered one.
2.2.2 Hypertext

Alternatively, the comparison process can be based on users' browsing over a well-constructed information network. Here, the index terms and their pointers to the Document Space provided in the traditional data structure become less important, and therefore, invisible to the users. Instead, those documents which share the similar information are explicitly linked together to formulate a connected information web. In such a system, it is unnecessary to specify explicitly the users' information needs. These users browse the Document Space from one information component to another by following the links provided in the information web; they decide which information component is relevant to their information needs (SIMILAR in Figure 2.3). Thus, the problems discussed in the previous section are effectively avoided here.

![Diagram of hypertext]

Figure 2.3: Functional overview of hypertext.

This kind of information web, which consists of information components and the intra- and inter-component links, represents a highly structured information management model called hypertext. Figure 2.3 shows the functional view of hypertext in an IR environment. Compared with Figure 2.2, the direct mappings from index terms to documents provided by the traditional data structure become less important in hypertext, and therefore, are often expressed implicitly. Here, the emphasis is on the documents which share the same (or logical development of) concepts. These
documents are linked together to formulate a connected information network, in which the user could browse from one document to another.

The configuration of hypertext, expressed as anchors (information) and links (relationship between different pieces of information), can not only represent information components (concepts), but more importantly, can represent the relationships among different information components (concepts). This new data structure, therefore, represents a higher level approach than that used in traditional IR. Hypertext opens exciting new possibilities for using the computer as a tool for thinking and communication, as well as for information organization, management and retrieval. The essence of hypertext is its associative structure, which follows human associative thought and memory. Support for hypertext and related systems is motivated by a belief that hypertext can provide more efficient and personalized access to text by complementing the global search techniques of traditional information retrieval models with local navigation based on meaningful intra- and inter-document connections. As a model, hypertext provides at least the following potential:

1. Hypertext opens a new way for thinking and information creation.
   - Hypertext offers machine support to augment human associative thinking and memory;
   - Hypertext presents a new environment for authoring and editing;
   - Hypertext encourages a collaborative working environment.

2. Hypertext opens a new way for information organization and management.
   - Hypertext creates a new environment for version management;
   - Hypertext supports divergent views;
   - Hypertext presents a new model for personal information management.

3. Hypertext opens a new way to retrieve information.
   - Hypertext offers an opportunity for both linear and non-linear reading;
   - This associative structure provides extra information about the relationships between concepts and between information components;
   - The user plays a more active role in the hypertext environment.

Hypertext and its applications in a massive and global information environment is the main topic of this dissertation. It is first reviewed in more detail in Chapter 3; a new hypertext model is then developed in the following chapters.
Chapter 3

Hypertext

This chapter presents a more detailed analysis of hypertext: its definition, its history, the models used in present hypertext systems and their characteristics. The emphasis is on investigation of the problems existing with the present hypertext systems, as well as the prediction of the next generation of hypertext. Based on the analysis, a soft-link hypertext model is proposed at the end of the chapter.

3.1 Definition

Hypertext systems provide the information users with the ability to follow their thinking associations and browse interactively in an Information Space via intra- and inter-document links to search for the information they needs. To be a hypertext system, one organization must have the following three fundamental elements [37]:

1. A database of information components;
2. A link structure which connects those information components;
3. Tools for creating and manipulating this combination of information components and the link structure.

3.2 History

In 1945, President Roosevelt's science advisor Vannevar Bush first proposed the idea of hypertext in his famous article As We May Think [27]. In the article, he described memex, a tool that provides access to a large collection of microfilm and mechanisms to make links between any two pieces of information in the system. What distinguished Bush's concept from other forms of data storage was its associative structure that followed closely the structure of human memory: "The human mind ... operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain" [27]. Although memex was never actually developed because of the technology limitations at that time, Bush's prediction of "selection
by association, rather than by index" has since driven generations of researchers to explore this arena.

### 3.2.1 The First Generation Hypertext

The first serious attempt to build a memex did not take place until 20 years after Bush's description. In the 1960s, Engelbart at the Stanford Research Institute was influenced by Bush's article; his team developed a computer system which embodied many of Bush's ideas. The system evolved over the years into a system called Augment [37]. Like other early hypertext systems, Augment emphasized three aspects: a database of linear text, view filters which selected information from this database, and views which structured the display of the information for the terminal. The idea behind the system is that the user and the computer were dynamically changing components in a symbiosis which had the effect of amplifying the native intelligence of the human user. Among the innovative features of Augment were its multi-user capability and multiple windowing, as well as two peripheral command devices to enhance the keyboard - the chord keys and the mouse.

Meanwhile, another attempt to build up a hypertext system was conducted by Ted Nelson. Again, Nelson's system (called Xanadu [37]) consisted of three parts: a database, a network of links and a front end. However, there the emphasis was on creating a unified literary environment on a global scale, instead of a mere enhancement of the user's personal information locally. Therefore, Xanadu's design makes a strong separation between its crude front end and the well-developed database server.

Augment, Xanadu and other first generation hypertext systems such as FRESS, ZOG, etc. represent the first step on the way to the realization of "selection by association, rather than by index". They were based on the most primitive model of hypertext: a database of documents, a logical network amongst different sections of a document or amongst different documents, and an interface which supports browsing across both the logical and physical networks. Because of the limitation of technologies involved in building up hypertext systems at that time, the scales of those systems, as well as the trials of "hypertext" as an idea, were very limited.
3.2.2 The Second Generation of Hypertext

In the 20 years since Engelbart's demonstration, both interest in, and activity regarding the development of, hypertext grew steadily until the 1980s since which enthusiasm for hypertext has accelerated sharply. The second generation of hypertext is identified as follows:

1. The formal theoretical studies of hypertext models and their representations have started [81] [82] [69] [150] [155] [38]; the research topics covered all elements of hypertext, with the emphasis on link structure [113] and hypertext languages;

2. Hypertext has been systematically tested and compared with other computer technologies [23] [117] [166] [112] [97];

3. The trials and experiments of hypertext systems have helped to extend the idea of "hypertext", as well as those regarding its fundamental elements [52] [37] [151];

4. Many hypertext systems were implemented and have been put into use; they include Intermedia [40] [78] [167], HAM [29] [30], KMS [4] [5], I²R [43], Hyperties [108], Guide [21], Neptune, Textnet [157], Issue-Based Information Systems (IBIS) [38] and etc.

5. Some hypertext systems were commercialized. The most noticeable projects include Xerox PARC's NoteCards [80] and Apple's HyperCard. These systems have done a tremendous public promotion for hypertext and hypertext systems.

6. The Document Space in those systems become increasingly large, the link structure increasingly complicated and the front ends increasing sophisticated.

7. Hypertext has been extensively applied in manual packages, distant learning [92], electronic publishing and collaborative working environments;

8. A series of problems have been found with the practice of hypertext [80]. The most severe ones include the generation of the hypertext link structures and the auxiliary tools needed to avoid the user getting lost in a hypertext Information Space. This point is discussed in more detail in Section 3.4.2;

9. The interest in, and demand for further development of, hypertext systems have been stimulated.
3.2.3 The Third Generation of Hypertext

The development of hypertext is now at the crossroads of the third generation of systems. The starting point of theory and practice of the latest generation is symbolized by Halasz’s famous paper [80]. The following attributes of the new generation represents the theoretical development and practice of hypertext for the foreseeable future.

1. Large-scale Hypertext;

The combination of hypertext, multimedia and the latest developments in global communication networks has made systems like World-Wide Web [11], Gopher [76], Hytelnet [140], etc., a reality. Although many new challenges lie ahead, the theory and practice of hypertext have taken another important step towards the vision of a massive and global information network.

2. Open and Reconfigurable Hypertext;

Hypertext systems should not be limited to closed infrastructure. Instead, the design and practice of hypertext should be open and reconfigurable, so that it can be easily extended to new information services and integrated with the diverse existing applications [88]. Open and reconfigurable hypertext is being experimented in pioneering systems including Microcosm [88] [51], Sun’s Link Service [122], VNS [141] and Hyperform [163].

3. Automatic and Reconfigurable Hypertext;

To reach the goal of a large-scale information network, the consistent, systematic and automatic generation of the link structure is crucial. Some investigations on the automatic generation of hypertext structure have been reported by Rada & Diaper [23] [125] [126].

4. Searchable Hypertext.

In some sense, hypertext stands for a navigational access to the information. The ability to browse around a network by following the link structure from an information component to another is a defining feature of hypertext. It is precisely this ability that makes hypertext a powerful tool for managing loosely structured information. However, the experience from the previous generations of hypertext suggests that navigational access by itself is insufficient. Effective access to information stored in an Information Space requires both query-based global access and browsing-based local access [80] [68]. Nevertheless, recent
theoretical studies and the practice of searchable hypertext have been limited to the following aspects:

- Theoretical analysis [68];
- Query-based access to a starting point for local browsing;
- Query-based access in a highly-structured hypertext environment;
- Query-based access in semi-searchable hypertext [151].

Labelled by many as “yet to be fully explored” [37] [80] [52], searchable hypertext represents a very important feature of the new generation of hypertext.

5. Intelligent Hypertext;

Whilst users are able to browse in many hypertext systems, they are still dogged by worries such as getting lost in the information web, or unable to find the required information. Consequently, their browsing is still constrained by many factors. Attempts to overcome these difficulties are being made by applying other related studies like artificial intelligence (AI), neural networks, psychology, probability theory, etc. in hypertext, with the following aims:

- To understand the information users and their information needs better;
- To create intelligent systems with some domain knowledge and reasoning ability.

Savoy [139], Frisse [67] & Cousins [68], Croft & Turtle [159] [44], and Rada [127] are amongst the pioneers creating such intelligent information systems. Their researches are basically tackling these problems. Furthermore, there is a trend to integrate hypertext more closely with conventional IR methodologies. Labelled as “computed links”, various technologies developed in intelligent information retrieval systems may be well applied in hypertext [151] [83] [51].

6. Hypertext and Hypermedia.

Since late 1980s, many multimedia systems have been developed to provide an integrated environment for the creation, storage and presentation of a variety of media types (e.g. text, graphics, audio and video). The combination of multimedia and hypertext leads to a new notion called hypermedia. In a hypermedia system, the information database includes not only plain text, but various other media types; its link structure connects these homo-media and hetero-media nodes to formulate a multimedia web. Currently, hypermedia is being investigated in a number of systems including Elastic Charles [118],
Microcosm [51] [83] [88]. This should further extend the user acceptance and application of original design.

3.3 Why Hypertext?

Hypertext is a computer-based medium for thinking and communication [37]. Its essence is the machine-supported anchors and links which are able to represent the more complicated concepts and the relationships among concepts; as a result, it share greater similarity with human mental associations. Hypertext makes the communications between machine and human beings easier. Hypertext can be compared and contrasted with other information media as follows:

1. Hypertext provides a better mechanism for delivering information than traditional documents.

Although traditional linear literature is also organized hierarchically and interlinked richly for nonlinear reading with cross-references, bibliographic references, dictionaries, encyclopedias, indexes, glossaries, etc, the machine-supported hypertext provides a superior model for the following reasons:

- It makes searching and following those referential trails easier and faster;
- It supports more dimensions of movement inside the Information Space, not only the linear ones like forward and backward, but also the nonlinear movements among information components; therefore, it provides greater flexibility and convenience;
- It facilitates the access to large, distributed and diversified complex information sources;
- It provides possibilities for using machine intelligence as an extension of human intelligence for learning, remembering and reasoning;
- Most importantly, as machine-supported hypertext links enhance the user's mobility in an Information Space and make these movements closer to the speed of human thinking associations, they provide tremendous help in keeping those associations continuous, or even help in stimulating new thinking associations.
2. Hypertext offers new possibilities for authors and editors, for collaboration and for dynamic revising and updating.

Whereas word processors provide good tools for word- and sentence-level authoring, hypertext goes further in helping the human being to explore concepts, to structure ideas and to organize the order of presentation.

3. Hypertext provides a more suitable model for information retrieval than general databases or database management systems (DBMS).

Although a database is a fundamental element of hypertext, and hypertext shares some similarities with database models (especially the recent ones like relational databases [107] [149], object-oriented databases [102] [23] and semantic databases [90]), they emphasize different aspects of information management and retrieval; thus, they are suitable for different applications. Whilst these differences are listed in [63], they can be summarized as follows:

- Hypertext and DBMS have different goals. Whereas DBMS is mainly for the manipulation of homogeneous data and logical operations of these data, hypertext allows the integration of heterogeneous information, as well as the exploration and browsing of the information;
- They have different emphases. Whilst a database or DBMS puts the priority on logical form over content, a hypertext system emphasizes the contents rather than the form;
- Although a database, or a DBMS, can totally ignore its users and operate on the basis of as-a-matter-of-fact, the essence of hypertext is to follow and mimic human associations; thus, the user's cognitive and psychological attributes are very important in hypertext applications;
- A database management system (DBMS) lacks the single coherent interface with stored information which is the hallmark of hypertext [37].

4. Even though most text processing tools share some similarities with hypertext systems, hypertext provides a superior model for information retrieval. This is because, as well as supporting hierarchical databases and sophisticated interfaces for viewing different components in the database, it also supports cross-references between components.
3.4 Models of Hypertext

3.4.1 Models of Hypertext

The study of hypertext models is still an undeveloped area. The three main models found in the literature are the Dexter model [81, 82], gIBIS [38] and the Trellis model [150]. Each of them clarifies certain aspects of the topic [69]. These three models are briefly illustrated as follows:

1. The Dexter model represents the most popularly used model of hypertext systems for information retrieval. It is based on the three-element hypertext principle and provides a simple and clear model. The model is composed of three separate layers and two mechanisms (Figure 3.1):

   - The Within-Component Layer, which provides a mechanism to define the composition of each type of component (e.g., text, table, equation);
   - The Anchor Mechanism, for communications between the Storage Layer and the Within-Component Layer;
   - The Storage Layer, which defines how text components and links can be related to form hypertext networks;
   - Presentation Specifications, for communications between Runtime Layer and the Storage Layer;

![Figure 3.1: The three layers of the Dexter model.](image-url)

...
• The Runtime Layer, which provides a mechanism to define how hypertext components and links will be displayed and manipulated.

As shown in Figure 3.1, the Dexter model presents the minimal features of a hypertext system, and emphasizes a Storage Layer where the fundamental features of hypertext are represented. Internal representation of information components is performed by the Within-Component Layer and mediated by Anchoring Mechanisms. Presentation is carried out by a Runtime Layer in conjunction with Presentation Specifications. The Dexter model is discussed in more detail in Section 8.2.1.

2. gIBIS was developed by Conklin and Begeman in 1989, based on the Issue-Based Information System (IBIS) [38]. The model defines a three-layer tree structure, with nodes of tree (i.e., issues, positions and arguments) as information components and arcs as the semantic relationships. As the authors put it, "Although gIBIS is a relatively weak form of hypertext, it has shown itself to be useful both in terms of structuring and preserving a complex line of reasoning and in terms of supporting an explicit rhetorical model." The link semantics imposed by gIBIS complement the structural features defined by the Dexter model and provide a framework for the semantics of associations between hypertext components and a primitive reasoning mechanism.

3. Trellis is based on Petri nets [150]. Places represent possible actions such as "display a component". Attributes of places can represent information about component and link parameters. Preconditions represent the necessary conditions to transfer from one state to another.

### 3.4.2 Problems Existing in Present Hypertext Systems.

Seven main problems existing in hypertext systems are listed as follows:

1. There is a lack of a proper abstract data structure for the hypertext associations.

   Until now, no formal data structures have been proposed for the systematic management of the information web (anchors and links) used in hypertext. This hinders further advances in the area. As suggested by Thuring [153]: "Clearly, it is not sufficient to merely link the nodes, since a link only indicates the existence of a relation without specifying its semantics". This makes hypertext confusing for both authors and information users.
The lack of abstract data structure reflects the fact that hypertext has not really been well understood. This is demonstrated more clearly in Figure 3.2. Here, information components are linked with one another to formulate a data structure for information organization and presentation. Nevertheless, the reasons that such a link is generated — different information components sharing similar concept(s), or logical development of concept(s) — are generally ignored. As a result, no systematic rules exist to govern why two components should be linked, let alone to explain it clearly to the user.

2. Difficulties in hyperization.

One direct result of the lack of a proper data structure is that few tools exist for systematically generating the hypertext links, either automatically or semi-automatically; hence, it is difficult to build up the information web for hypertext, or to convert ordinary texts to hypertext (and vice versa). This problem involves the following three separate but interrelated activities:

- Identification and generation of the individual components as potential hypertext nodes in the Document Space;
- Identification of concept(s) as potential hypertext anchor(s);
- Identification of the interrelationships between the separate components, or concepts, to formulate hypertext link structure.
At the moment, most of the hyperization is done manually; this limits the scale of information webs that can be provided, as well as incurring a high cost in human effort involved in building up the network. As one example, the hyperization of 8 papers about hypertext in the July 1988 issue of Communications of the Association of Computing Machinery took 12/person-week to complete the work [23]. Furthermore, these manual settings do not always represent the best presentation of the Information Space.

In order to build up a massive and global network of information with a large number of intra- and inter-document links, an automatic hyperization mechanism is necessary for two reasons:

- To help authors and editors to set up associations efficiently between information components, or between concepts;
- To automatically generate the data structure for hypertext.

So far, very limited work has been done on the automation of hypertext [70] [23]. Furthermore, it is more successful dealing with highly-structured texts like dictionaries, manuals, directories, catalogues and electronic yellow pages [23]. Hyperization of lowly-structured texts is still an open question (Chapter 5).

3. Structure of the hypertext network.

Many aspects of the data structure needed for hypertext are still unknown. For example, in what form the information components in the Document Space should be stored and represented, how many links a hypertext system should facilitate and how these links should be be organized and presented, etc..

4. Searching for facts versus browsing.

Browsing is an effective way to formulate a description of an information need and mimics the information user's thinking associations; however, it is limited by its local-access methodology, whilst global indexes always provide instant and efficient access to the Information Space. As Frisse puts it [68], "Because indexes are in one sense a set of 'pre-compiled' links, they facilitate access to a needed information unit without the need for traversal through many intermediate information units and the time necessary to find indexed information does not increase significantly as the size of the document space increases."

Which of them presents a better information retrieval pattern may depend on the nature of the users' thinking associations at a specific time. For example, when this association is a continuous, local and blurred process, the user may
feel comfortable in browsing for information with recognition playing an important role. Nevertheless, when this association process is a jumpy, discontinuous and precise process, a global index could speed up the search and retrieval process so as to present a better information retrieval model. Here, recall can play a more important role.

Therefore, a good solution may be a proper combination of both methods in one information retrieval model. As suggested in the literature [68] [154] [80], many researchers realize the problem, but little has been done to resolve it.

5. User (dis)orientation;

Along with the power to organize information much more diversely comes the problem of users needing to know:

- Their locations in the network;
- Which link(s) to follow to improve the probability of locating the desired information.

Without explicit and clear representation of their data structures, hypertext systems usually do not provide adequate help for their users. This quite often leads to the so-called “getting lost” problem in the hypertext literature. Obviously, a user might also have disorientation problems with traditional linear documents; but there, the options are only limited to forward or backward search. In hypertext, as more degrees of freedom and more dimensions of movement are offered, the potential for a user to get lost, or disoriented, becomes greater, especially when the network of knowledge becomes massive and global.

Another aspect is that finding the location in a database should not be as difficult as knowing where to go next. Whereas the former can be resolved satisfactorily with the existing computer technology, researchers still can not find a good solution for on-line navigation. Many graphical tools have been developed to help browsing or getting oriented; they include Fisheye, Map, History graphs, Twin cards, etc. However, these tools can be efficient only in some limited applications, or alleviate the problem only partially [116]. A thorough solution may still be the provision of a more transparent data structure for the associations. In this way, not only are the information components connected, but also the mechanism and reason behind such a connection [concept(s) and information shared by different documents], can be shown to the user. As a result, the users can overcome their passive roles in the hypertext environment,
and take more active parts in information retrieval and browsing. This reflects the essence of hypertext.

6. Cognitive overhead;

Hypertext tends to present the readers with a large number of choices as to which links to follow and which to leave alone. These choices engender a certain overhead in decision making; an overhead that is absent when the author has already made many of these choices for you. In other words, an additional effort and concentration is needed to maintain several tasks or trails at one time. This aspect of hypertext has advantages when the richness is needed, and drawbacks when it is not [37].

7. Learning and improvement.

Again, without a thorough understanding of hypertext and its data structure, learning and improvement on hypertext structures becomes almost impossible. Conklin addressed this problem in 1987 [37] and Halasz put it as one of the main issues for the future development of hypertext models and systems [80]. However, no solutions have yet been proposed.

3.5 Soft-Link Hypertext

The conclusion drawn from the above discussions is that the major obstacle to further development of hypertext lies on the implicit nature of its link structure. This implication disregard the semantic and logical relationship between two linked information components (concepts), limit the expressibility of the link structure and leads to the confusion to both the information authors and users. Should this relationship be made explicit, several problems discussed in Section 3.4 could be solved. It should be noted that making the hypertext link structure explicit would not be too difficult as the global index, based on which the hyperization happens, already exists and has been fairly-well studied.
Investigation of a new hypertext model, with its explicit link structure based on a special mechanism called *conceptual index*, constitutes the main topic of this research.

Figure 3.3 represents a different approach to browsing than the conventional hard-link hypertext model shown in Figure 2.3. In the conventional hard-link hypertext environment, browsing from Document 1 to Document 2 follows a pre-set hypertext link (Figure 2.3); therefore, problems exist and the data structure confuses the user. In the soft-link hypertext model, after a user clicks a hypertext source anchor in Document 1, he/she first goes to the Index Space where an explicit link structure is displayed. There, a proper link can be chosen and followed, according to the user's information needs. Finally, the user goes to Document 2 by following the hypertext link and the pointer provided by the conceptual index (Figure 3.3).

The difference between traditional search and retrieval, traditional hard-link hypertext and soft-link hypertext is also illustrated in Figure 3.4. In each of these models, implicit operations are represented in dashed boxes. For example, in the traditional hard-link hypertext model, the conventional indexing and searching operations are usually disregarded. Instead, documents are linked one another in hyperization,
whilst the information user explores the Information Space by following these links from one document to another during his/her browsing procedure. Nonetheless, a closer study reveals that, even in the model, both indexing and searching are actually applied. They are embedded in the hyperization process. As Frisse pointed out [68], the conventional index structure in fact accommodates a set of "pre-compiled" links. The human who generates the link structure in the traditional hypertext environment has actually gone through the indexing and searching processes either consciously or subconsciously.

<table>
<thead>
<tr>
<th>1. Traditional search and retrieval model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing: Document ---&gt; Keyword</td>
</tr>
<tr>
<td>Searching: Keyword ---&gt; Document</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Traditional hard-link hypertext model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing: Document ---&gt; Keyword</td>
</tr>
<tr>
<td>Searching: Keyword ---&gt; Document</td>
</tr>
<tr>
<td>Hyperizing: Document ---&gt; Document</td>
</tr>
<tr>
<td>Browsing: Document ---&gt; Document</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Soft-link hypertext model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing: Document ---&gt; Keyword</td>
</tr>
<tr>
<td>Searching: Keyword ---&gt; Document</td>
</tr>
<tr>
<td>Hyperizing: Document ---&gt; Document</td>
</tr>
<tr>
<td>Browsing: Document ---&gt; Keyword ---&gt; Document</td>
</tr>
</tbody>
</table>

Figure 3.4: The difference between three different IR models.
The soft-link hypertext model has the following potential advantages:

1. When the Index Space plays an important part as the *link structure* in the soft-link hypertext model, the tedious and expensive manual hyperization may become redundant. Automatic indexing and hyperization are possible;

2. If the users need to choose which link(s) to follow, the whole structure must be explicit for them. Here, the information users are more actively involved in the information retrieval and browsing procedure; a hypertext link is followed because the user wants it, not because the author wants it;

![Diagram of the conceptual index in the soft-link hypertext model.](image)

**Figure 3.5**: The conceptual index in the soft-link hypertext model.

3. The core of the soft-link hypertext model lies in a special data structure, called the *conceptual index*, adopted in the Index Space (Figure 3.5). Traditionally, the global index provides the mappings from index terms to documents and facilitates global and immediate access to any index terms in an Information Space (Section 2.2.1). In the soft-link hypertext model, the conceptual index is composed of a semantic net built upon the conventional index mechanism; therefore, it not only inherits the characteristics of the global approach, but more importantly, supports local associations among index terms (concepts).
With such a structure, the real beneficiary is the user, whose mobility in the Information Space should increase and the problem of getting lost should be alleviated;

4. The soft-link hypertext also breaks the limitation of the conventional one-to-one hypertext structure and facilitates one-to-many associations from the source anchor to the destination. For example, in the conventional hard-link hypertext environment (Figure 3.2), invoking hypertext mechanism in Document 1 usually means that the user is transferred to another specified document, Document 2. There, the assumption of hypertext is from one document to another, e.g. Document 1 to Document 2, a one-to-one structure. Nevertheless, in the soft-link hypertext (Figure 3.5), this one-to-one structure no longer exists. A user can be transferred to many different destination document (Document 2, Document 3, ..., Document j) from Document 1, depending on which soft link is chosen;

5. A sophisticated semantic net can be adopted in the Index Space to support various kinds of associations and relationships. As a result, a much richer set of entities can be expressed. This leads to many positive effects for information retrieval applications including hypertext;

6. Machine learning mechanisms can be used on the conceptual index to facilitate self-adjustment and evolvement of the structure, and improve its usability; they can also provide some extra information for the information user, and make information retrieval more efficient.

Full development of the soft-link hypertext model, which includes its theory and implementation, is presented in detail in the following chapters. Chapter 5 introduces the special conceptual index structure used in the soft-link hypertext model: on the one hand, the conceptual index still provides the mapping from each individual concept to the related documents, and therefore, supports the normal Boolean search model; however, on the other hand, the semantic net adopted in the conceptual index constitutes the relationship among concepts, and therefore, supports local hypertext associations. A set of non-syntactic & non-semantic formulation rules are used to formulate the conceptual index. These rules are also presented in Chapter 5. The methodology and the mechanisms of the soft-link hypertext model are developed in Chapters 6 and 7, whilst a version of its implementation is presented in Chapter 8. Finally, the model and its implementation are evaluated in Chapter 9.
Chapter 4

Conceptual Graph

A conceptual graph is defined in this chapter, with which the conceptual index used in the soft-link hypertext model can be represented as a graph, \( G = (V_C, V_R) \). In this way, the relevant notions in graph theory can be well adopted to describe the state-space of the Information Space, as well as its applications in information retrieval. The nodes in the conceptual graph represent different information components (or anchors), the arcs represent the relationships between these components (hypertext links), and finally, the paths represent the information user's browsing trails in an Information Space. Although the conceptual graph might not seem to be of much help at this stage, it provides a theoretical base and formal notation for further presentations in following chapters.

4.1 Introduction

The theory of conceptual structure was originally used in the semantic network of knowledge-based systems. It forms a knowledge representation language based on linguistics, psychology and philosophy [145], in which concept nodes are used to represent entities, attributes, states and events, whilst relation nodes show how the concepts are interconnected.

The immediate advantage of adoption of the conceptual graph is that all notions in the graph theory can be easily introduced into the dissertation. With its elements endowed with some semantic meaning, the conceptual graph becomes very useful in presenting the soft-link hypertext model and its applications in information retrieval. This presentation not only includes the state-space of the model, but also covers the dynamic operations and all related events happening in the model. As a result, all IR practices can be explained by the related operations in the conceptual graph. For example, formulation of the conceptual index can be explained as the addition of nodes and arcs, removing information components as deletion of nodes and arcs, and information retrieval and browsing as paths, or trails, on the conceptual graph.
4.2 Some Definitions

In this dissertation, I adopt three conventional terms defined by Sowa [145]. They are *concept*, *conceptual relation* and *conceptual graph*.

Definition 4.1 In the theory of conceptual structure, basic primitives are called *concepts* [144].

Concepts represent the smallest information unit in an Information Space. Represented in a language, concepts can be either expressed by a word, or the combination of words (phrases). For example, in Figure 4.1, Hypertext is a concept, Definition of Hypertext is another concept.

![Figure 4.1: A conceptual graph.](image)

Definition 4.2 Connections between concepts are represented by *conceptual relations* [144].

Definition 4.3 A *conceptual graph*, $G = (V_C, V_R)$, is a finite and directed graph:

- $V_C$ represents the set of *concept*;
- $V_R$ represents the set of *conceptual relation*:
  * $V_R \subseteq V_C \times V_C$;
  * $V_R = \{V_{Rij} = (V_{Ci}, V_{Cj}) \mid V_{Ci} \in V_C, V_{Cj} \in V_C\}$;
  * $V_{Cj} = V_{Rij}(V_{Ci})$. 
Figure 4.1 shows a conceptual graph with nine vertexes and nine arcs. Each vertex represents a concept. The arcs represent separately the conceptual relations definition, model, advantage, database, history, prototypes, extension and application. It can be seen that a single concept by itself may form a conceptual graph, but every conceptual relation must be linked at least to two concepts. The conceptual graph provides a graceful notation to represent the data structure of any Information Space. If all the concepts accommodated in an Information Space is noted as a set Concept, it should be a part (i.e., a subset) of the concept exists in the universe (which is noted as CONCEPT), i.e., Concept ⊆ CONCEPT.

Furthermore, in order to make the further presentations easier, another three terms are defined: immediate successor, immediate predecessor and conceptual path.

**Definition 4.4** Vertex $V_{c_i}$ is called a immediate successor of vertex $V_{c_0}$ if there exists an arc $(V_{c_i}, V_{c_0}) \in \mathcal{V}_R$. The set of all immediate successors of $V_{c_0}$ is denoted by

$$\Gamma^+_G(V_{c_0}).$$

**Definition 4.5** Vertex $V_{c_i}$ is called a immediate predecessor of vertex $V_{c_0}$ if there exists an arc $(V_{c_0}, V_{c_i}) \in \mathcal{V}_R$. The set of all immediate predecessors of vertex $V_{c_0}$ is denoted by

$$\Gamma^-_G(V_{c_0}).$$

- The set of all immediate neighbours of $V_{c_0}$ is denoted by

$$\Gamma_G(V_{c_0}) = \Gamma^+_G(V_{c_0}) \cup \Gamma^-_G(V_{c_0}).$$

- The degree of vertex $V_{c_0}$ is the number of arcs with $V_{c_0}$ as an endpoint. The degree of $V_{c_0}$ is denoted by $d_G(i) = d^+_G(i) + d^-_G(i)$.

**Definition 4.6** A sequence of concept pairs, $(V_{c_0}, V_{c_1}), (V_{c_1}, V_{c_2}), ..., (V_{c_{n-1}}, V_{c_n})$ is called a conceptual path, noted as $P = \{V_{c_0}, V_{c_1}, ..., V_{c_n}\}$, when

- $V_{c_i} \in \mathcal{V}_C$, where $i = 0, 1, ..., n-1$;
- $V_{c_{i+1}} = V_{R(i+1)}(V_{c_i})$, where $V_{R(i+1)} \in \mathcal{V}_R$, $i = 0, 1, ..., n-1$.
- $n$ is the length of the conceptual path;
4.3 Conceptual Structure in Soft-link Hypertext

As shown in Section 3.5, a special data structure, called the conceptual index, is used in the Index Space in the soft-link hypertext model. It consists of a semantic net built upon the conventional global index structure (mapping from the concept to the document); as a result, the conceptual index combines the conventional global index mechanism with the related semantic relationships among different index terms. Therefore, it supports both global access to, and local associations among, the concepts. With the conceptual structure defined in the last section, the conceptual index can be represented as a conceptual graph, \( G = (V_c, V_r) \) (Figure 4.2).

1. Each vertex in the graph represents an individual concept in the conceptual index. Each vertex has an attribute associated with it, which provides a list of pointers to the related documents. The attribute constitutes the mapping from concepts (index terms) to documents. Therefore, the conceptual subgraph, \( G' = (V_c, \text{NULL}) \), by itself accommodates the conventional index structure; For example, when the user expresses the information need as a concept (Hypermedia), all the documents including the concept, Hypermedia, can be retrieved by following the pointers associated with the related attribute.

![Figure 4.2: A conceptual index in the soft-link hypertext model.](image-url)
2. Each conceptual path with length 1 (concept → conceptual relation → concept) facilitates a local association between a pair of concepts. In fact, this association represents a hypertext structure (source anchor → hypertext link → destination anchor). As a result, it is easy to see that the conceptual index constitutes the link structure suitable for hypertext (Figure 4.3).

For example, when the user invokes the soft-link hypertext by clicking Hyper-text in Document 1, the corresponding concept (Hypertext) in the conceptual index represents the source anchor and serves as a starting point of a potential hypertext association. Here, each of the arcs from the concept Hypertext to its immediate successors constitutes a soft hypertext link. By choosing one of these soft links, e.g., (Hypertext, Models of hypertext), the user could be traversed to another concept (Models of hypertext in this case) and to Document 2 in which the destination anchor Models of hypertext is included.

Figure 4.3: The correspondence of different layers.
Thus, it is easy to see that such a conceptual index used in the soft-link hypertext model has at least the following advantages:

1. If a user's browsing during a hypertext session is noted as a set of conceptual paths, $PATH^* = < P_1^*, P_2^*, ..., P_i^* >$, $PATH^*$ should be the copy of that expressed implicitly in the user's mind and reflect his/her real information needs. $PATH^*$ is used in Chapter 7 for inferencing the user's next move;

2. When the conceptual index is adopted in the Information Space, the hypertext structure becomes explicit and flexible. Furthermore, the tedious and expensive manual hyperization can be saved (Figure 3.4). Here, hyperization is turned into the formulation of the semantic net needed for the conceptual index; thus, automatic hyperization process becomes possible;

3. With the soft-link hypertext model, the operation of information retrieval process becomes:

   Document i → Conceptual index → Document j → ...

   Here, the focus of IR operations is shifted to the Index Space. The organization and manipulation of the Information Space becomes the operations on the conceptual index in the Index Space.

The questions that now arise concern how such a conceptual index can be automatically, or semi-automatically, generated; what kinds of intelligence can be installed on such a data structure; finally, how it can be used to help in information search and retrieval. These are the questions to be addressed in the following chapters.

First, in Chapter 5, an automatic index formulation mechanism is introduced to build up the conceptual index. Based on the conventional full-text indexing methodologies and a set of non-syntactic & non-semantic formulation rules, a global index and the related semantic net are automatically generated. Chapter 6 defines a connectionist network, based on which the individual mechanisms in the soft-link hypertext model are built. The dynamic operations and applications of the soft-link hypertext model are governed by these mechanisms. These mechanisms are developed, in Chapter 7, to constitute the soft-link hypertext model. The model, as well as a version of its implementation, is presented in Chapter 8. Finally, in Chapter 9, the soft-link hypertext model and an implemented system are evaluated.
Chapter 5
The Conceptual Index and Its Formulation

Chapter 5 describes the conceptual index and its automatic formulation. It first reviews the diverse data structures and formulation methodologies used in different IR environments. Based on such an analysis, an automatic formulation mechanism is developed to generate the conceptual index in the soft-link hypertext model. Such a generation process includes two different, but related tasks: developing a global index, and building a semantic net on it. The formulation mechanism adopted in the soft-link hypertext model is based on the conventional full-text indexing methodology and a set of non-syntactic & non-semantic formulation rules. An implementation of the formulation mechanism, as well as an experiment designed to assess the mechanism, is also described briefly in the chapter.

5.1 Global Index and Indexing

As discussed in Chapter 2, an index is a data structure used in information systems to speed up the searching process, whilst indexing deals with the mapping from the Document Space to such a data structure. The idea is to formulate some content identifiers outside the Document Space, which could provide a more systematic view of all information components. In this way, searching can be made more efficient by comparing users' queries with these identifiers, instead of with information components one by one.

In a traditional IR environment, those identifiers, known as keywords or index terms, are usually collected and stored alphabetically in special files called inverted files. There, searching is to locate specific index terms in an inverted file. This ensures rapid and straightforward access to any index terms in the Information Space. Furthermore, different software tools could be used to speed up this locating process. Inverted files are popularly used in all Boolean and extended Boolean models. Because of its global accessing methodology, such a data structure is called the global index in this dissertation. The attributes of the global index can be summarized as follows:
1. The global index facilitates one-step access to any needed index items (concepts) without the need for traversal through many intermediate ones. In another words, it always ensures fast and efficient access to any specific index terms in an Information Space;

2. The time necessary to find a specific index item does not increase significantly as the size of the Document Space increases;

3. The global index adopted in the traditional IR models represents a low-level data structure for information organization. This is mainly because its content identifiers can only express simple concepts, but are inadequate for composite concepts, nor relationships between concepts (Section 2.2.1). As a result, it is only suitable for some simple, bibliographic information-seeking processes (Section 8.4).

Figure 5.1: Resolving power of significant (medium-frequency) words.

Indexing for the global index is to choose a set of keywords, which are able to most efficiently identify different documents [134] (Figure 5.1). Meanwhile, a specific relationship among documents, measured by the number of index terms shared among different documents, can be represented with such a data structure. This is called clustering.
5.2 Local Hypertext Structure

Another data structure, based on identifiers known as anchors and links, is exclusively used for hypertext. Unlike the global approach, the mapping from the concept(s) to different documents become less important in hypertext; therefore, it is usually neglected. Instead, different documents (or different sections in one document) which share similar concepts are linked together to form an information web (Section 2.2.2). It can be seen that the hypertext link structure is more document-to-document, or concept-to-concept based; thus, it constitutes a local approach. In hypertext, searching becomes a browsing process from one information component to another via the link structure. Such a local data structure has the following advantages:

1. This local hypertext approach offers another data structure to organize the information, and represents some added knowledge to the Document Space:
   - Anchors provide a richer set of expressions than the ordinary keywords and Boolean operators in representing the concept. They are able to cover both the simple concepts (keywords), and composite concepts (phrases);
   - Links represent the relationships between concepts. The mere presentation of a link by itself implies the existence of more information on the chosen concept, or some extension of it; therefore, the link provides an extra information in the Information Space. These extensions may include semantical, logical, or even conceptual expansion of the concepts.

2. The hypertext structure, emphasizing on various relationships and associations between concepts or documents, offers a higher-level structure which is more similar to that in human memory. It supports "selection by associations";

3. The hypertext structure supports local browsing which is more suitable for psychological recognition processes and continuous thinking associations;

4. The hypertext structure presents a more sophisticated extension of knowledge of the Document Space than the traditional data structure. It is, therefore, suitable for more sophisticated information-seeking processes (Chapter 8).

Formulation of the hypertext data structure requires the identification of anchors for proper concepts, and building up links among information components, or concepts. This is traditionally known as hyperization. However, until now, due to the inadequate natural language processing technologies, the automatic identification of these
composite concepts (anchors) and relationships between concepts (links) is still very
difficult, if not totally impossible. Hyperization is, hence, still an open question.

As seen in the literature [37], automatic hyperization has only been successfully ap­
plied in the following two special hypertext applicational environments:

1. **Logical document structure → hypertext**: When a document is organized in
such a way that each of its logical unit (document, chapter, section, para­
graph...) is self-contained, its logical structure can be adopted straightforwardly
as the hypertext structure. In this way, each logical unit is naturally fitted into
a unique information component on the subject. The title, headline, or even
the first couple of words of the component can be used as the content identifier;

2. **Highly-structured text → hypertext**: Automatic hyperization of a highly-
structured Information Space (e.g. directory, electronic yellow page, dictionary,
encyclopedia, catalogue and manual, etc.) is usually easy and straightforward,
because of its self-contained structure.

Besides these two exceptions, hyperization of ordinary lowly-structured documents
is always done manually; therefore, it is always very difficult and time-consuming.
One of the past experiences is to change 8 journal papers (Communications of
the Association of Computing Machinery, July 1988) into hypertext [23]. The
papers were written independently for journal publication, linearly with minimal
cross-referencing across papers. Using Hyperties, it took 12 person-weeks to complete
this hyperization process: the eight papers were first cut into 86 Hyperties articles,
38 figures and 120 references (low-structured text to high-structured text); then an
overview was added to each Hyperties article to provide a set of links to topics
(provision of hypertext link). As a conclusion, Brown said that “While some portions
of the conversion could be automated, it seems likely that conversion of a similar
collection of scientific papers would require a similar amount of overall effort” [23].

### 5.3 Relation between Global and Local Approaches

Although the two approaches discussed in last two sections represent two different
information organization and retrieval methodologies, and are used in different IR
models, they are strongly related to each other. The reasons can be listed as follows:
1. Both of them are data structures outside the Document Space.

2. None of these structures by themselves are of any value for the information user. They are only auxiliary tools used to present more information outside the Document Space so as to make information search and retrieval more efficient;

3. Global index supports global navigation; it is in one sense a set of "pre-compiled" links [68]. The fact that two separate documents are linked together in hypertext by itself implies that these documents are either about the similar concepts, or one is on the logical development of another (Figure 3.5);

4. They provide different accesses to a shared "target" (Figure 5.2) [68].

In fact, these two approaches are often proposed as complementary in the literature [80] [68]. Unfortunately, although some researchers have addressed the combination use of these two data structures, the applications are limited to the following practices:

1. The two approaches are treated as two separate, but complementary, data structures used in different IR systems, or in separate search and retrieval processes in one system, e.g., the global index for choosing a starting point and hypertext structure for local browsing;

2. These two mechanisms are only combined and used successfully in the highly-structured hypertext environment [67];
3. The two mechanisms are combined and used with the support of thesauri-styled semantic networks, like Bayesian network and belief network employed in some IR systems [139] [68].

In the next section, a conceptual index, which combines both global and local mechanisms, is proposed. An automatic formulation mechanism for the conceptual index is developed in Section 5.5.

### 5.4 Conceptual Index — Global + Local

Based on the discussions in last several sections, it can be seen that global and local approaches are two different, but complementary data structures. They represent different levels of information organization and have different emphases; they are suitable for different applicational environments. "The challenge is how to combine these two mechanisms properly inside one framework" [116].

As the first step to build up the soft-link hypertext model, a conceptual index (shown in Figure 5.3) is introduced in the Index Space. It combines these two approaches (global and local) in one IR infrastructure. The structure of the conceptual index is described in more detail as follows:
1. The essence of the conceptual index is a semantic net of concepts (Figure 5.4). Expressed with the conceptual structure defined in Chapter 4, the vertex of graph represents concept and the arc conceptual relation;

![Figure 5.4: A semantic net supports the local associations](image)

2. In the conceptual index, each concept has an attribute associated with it. The attributes provide a list of pointers to different documents in which the concept is included (Figure 5.5);

![Figure 5.5: Each concept has an attribute attached to it.](image)

3. Such a conceptual index provides two different approaches to access the Document Space. This can be explained as follows:
With its attribute, each individual concept in the conceptual index can provide the mapping from itself to the related information components (Figure 5.5). Thus, the conventional global index structure is supported by the conceptual index. Its index term is each individual concept; it serves exactly the same functions as inverted files do in the traditional IR environment. Boolean operations can be also applicable in the procedure;

Another approach, provided by concepts and conceptual relations, represents the mapping from concepts to concepts and then to information components (Figure 5.3). This concept-to-concept mechanism is very useful for the following reasons:

(a) It could be used to express expansions from simple concept(s) to complicated concept(s), or to phrases;
(b) It could be used for synomyic/isomorphic/semantic relationships. [Frisse [67], Savoy [139] and Croft & Turtle [44]'s applications of semantic network (belief network) in IR belong to this group];
(c) It could be used for more sophisticated conceptual relationships, such as Definition, Applications, etc. (Figure 5.4).

4. Whereas the global index facilitates global and rapid access to any concept in the Index Space, the local conceptual structure facilitates local "selection by association", which constitutes a higher-level structure and is suitable for more complicated IR tasks like hypertext (Figure 5.4);

5. Each association on the conceptual index is a potential hypertext link, a soft link, by invoking which an information user could be transferred from one concept to another, or from one information component to another (Figure 5.4);

6. With a conceptual index, both global and local structures are combined in one framework;

7. The formulation of such a conceptual index could be fully automatic:

   - Indexing methodology for the global index has been well studied [134];
   - Automatic formulation of the required semantic net is still very difficult, due to the present limitation of natural language processing technologies. However, an attempt is made in this research, which applies a set of non-syntactic & non-semantic formulation rules on each index term to formulate the concept-to-concept mapping between it and other related concepts. Hopefully, this mechanism is able to identify composite concepts
and conceptual relations to be used as anchors and links in hypertext. Furthermore, intelligent mechanisms used in the soft-link hypertext model are applied to improve the formulated conceptual index.

In next two sections, an automatic formulation mechanism designed to generate such a conceptual index structure is presented. An implementation of the mechanism, as well as the design of an experiment to test its feasibility and effectiveness, is described briefly in Section 5.6.

5.5 Formulation of the Conceptual Index

Formulation of the conceptual index structure in the soft-link hypertext model is an automatic process. It includes the following steps:

1. Based on the conventional full-text indexing methodology [134], all information components in the Document Space are analyzed, and a global index is firstly generated. Each of these index terms constitutes a concept;

2. The second step is to formulate a semantic net. This is to identify both the composite concepts and the relationships between concepts.

Whereas the global indexing is based on the observation that the frequency of occurrence of individual word types (that is, of distinct words) in natural language texts has something to do with the importance of these words for purposes of content representation [134], the formulation of the semantic net is based on the hypothesis that, if a proper stoplist is used, composite concepts, as well as the relationships between concepts, can be formulated by finding the closest (nearest) neighbour of each occurrence of a simple concept (keyword);

Traditionally, about 250 high-frequency function words are usually used during the global indexing as stop words. They comprise 40 to 50 percent of the text words, serve as poor discriminators and need to be eliminated [134]. However, the resolved index terms from the traditional indexing mechanism are still too general and extensive for conceptual relations. Here, a second and longer list of stop words is needed to limit the set of terms for conceptual relations (Figure 5.6).
Figure 5.6: Resolving power of significant words for the semantic net.

Figure 5.6 shows the resolving power of significant words for conceptual relations needed for the formulation of the semantic net. Compared with Figure 5.1, it can be seen that the emphasis of the semantic net formulation is different from that of the global index. Therefore, they have different requirements on the stoplist.

3. Next, for each index term on the global index, the following set of non-syntactic & non-semantic formulation rules is applied to find the nearest meaningful neighbour for each of its occurrences. The located neighbour constitutes the relationship, whilst the whole phrase represents a composite concept:

- The location of each hit of the index term in the Document Space is found;
- Its left and right immediate neighbours (words) are considered;
- If a stop word is met, the search continues on that side;
- If a punctuation (, ; :) is met, the word is chosen from the other side;
- Of the left-side and right-side options, the closer one is chosen;
- If both the left-side and right-side options are equally acceptable, use the left-side one;
- English morphology is applied on the chosen word to generate its normal form. Several well-known algorithms exist for the removal of word endings,
generally based on the use of a list of suffixes followed by the removal of the longest suffix matching any entry on the suffix list [134].

The chosen word is the sought conceptual relation and the whole phrase constitutes a composite concept. For example,

"...As reviewed elsewhere, rates of solute diffusion in organic polymer can vary over several orders of magnitude and will be a function of the nature and extent of polymer cross-linking..."  (Long-term Sorption of Halogenated Organic Chemicals by Aquifer Material.2. Intraparticle Diffusion. Environmental Science and Technology, Vol 25, No. 7, pp. 1237 - 1249)

The global index can provide a pointer, in the inverted file, from the term polymer to this document. During semantic net formulation process, the physical location of this sentence in the database is found first; then, both the left immediate neighbour (organic) and the right immediate neighbour (can) of the term polymer are examined. organic is identified as a valid term for conceptual relation, whereas can is not as it is in the stoplist. According to the above formulation rules, organic is chosen as the term to represent a conceptual relation and organic polymer as a whole represents a composite concept, and as a proper hypertext destination anchor.

4. Repeat 3, until all index terms on the global index have been considered;

5.6 Implementation & Experiments

The formulation mechanism presented in the Section 5.5 has been implemented in Tcl & Tk on SuperBook, as a part of the new soft-link hypertext system called the Enhanced SuperBook. The implementation is presented in more detail in Chapter 8. As the techniques needed for real implementation are beyond the interests of this research, the following steps are taken to simplify such an implementation process:

1. SuperBook fonts are not considered for any concepts (index terms);
2. Capitals are accepted for concepts (index terms);
3. A small program is used to apply some simple morphology rules, so that the following items are treated as one word:
   - Glossary word — polymer:
(a) polymer  
(b) polymer,  
(c) polymers  
(d) polymers:  

• Glossary word — (polymer) PS-I:  
  (a) polymer (PS-I)  
  (b) polymer, PS-I  
  (c) polymer (PS-I, Figure 1)  

4. If one of the two options is empty, use the other one instead;  
5. If both of the options are empty, use the keyword itself instead;  
6. All number (0-9, I, II, etc.) options are treated as stop words;  
7. All options beginning with a letter, ending, or including numbers are acceptable.  
8. When a search reaches the end of a line, it continues at the beginning of the next line. When a search reaches the beginning of a line, it goes backwards to the previous line and continues at the end of the line.  
9. For an index term which has immediate successor(s) in the semantic net, the attribute attached to it is deleted. The accumulated collection of attributes from all of its successors constitutes the global mapping from it to the related information component;  
10. For those index terms which have no immediate successors in the semantic net, their attributes remains in the global index mechanism.  

An experiment has been designed to test the efficiency and effectiveness of the automatic index formulation mechanism. Whilst it is discussed briefly here, the detailed presentation of experimental procedure and the results are given in Chapter 9.  

5.6.1 Experiment 1 — Index Formulation  

The purpose of the experiment is to test the effectiveness and efficiency of the automatic formulation mechanism for the conceptual index in the soft-link hypertext model. It includes:  

1. Testing and evaluating the time needed for formulating the conceptual index in the system.
2. Testing and evaluating the memory space needed for storing the conceptual index;

3. Testing the effectiveness of the soft links formulated;
   - Evaluation of the formulation process based on domain experts;
   - Evaluation of the formulation process based on traditional IR criteria:
     (a) Recall;
     (b) Precision.

4. Analyzing the experiment results;

5. Adjusting the non-syntactic & non-semantic formulation rules based on the experimental results.

Full presentation of the experiment, as well as its methodologies, results and the analyses of the results, is in Chapter 9.

5.7 Summary

In this chapter, a new and more sophisticated data structure has been introduced as an organization and presentation configuration for information retrieval. The conceptual index not only facilitates the features of conventional inverted files used in traditional IR models, but more importantly, it accommodates a conceptual structure which could be used for more complicated IR tasks such as hypertext. This conceptual structure, called soft links, presents a more formal configuration for the generation, management and manipulation of anchors and links in hypertext. Whereas an automatic conceptual index formulation mechanism used in the soft-link hypertext model has been described in the chapter, the resolved imperfect data structure could be further improved by the intelligent mechanisms in the model. These mechanisms are developed in the following chapters.
Chapter 6
Spreading Activation and Learning

This chapter introduces a connectionist network, based on which the rules governing all dynamic operations and applications of the soft-link hypertext model are defined. A review of the connectionist network and its applications in IR are given first. A connectionist network is then defined, in Section 6.2, to represent the conceptual index. Whilst the network by itself is enough for the state-space of the soft-link hypertext model, its forward activation spreading and backward learning provide very powerful mechanisms to control the dynamic operations of the model. Here, the activation spreading is interpreted as a reasoning and navigation process; the backward learning is used to self-adjust and improve the conceptual index. Activation and learning rules are also discussed in the chapter.

6.1 Connectionist Network

Connectionist networks represent information as a network of weighted, interconnected nodes. They have become popular in last few decades in information processing, cognitive science, artificial intelligence, and information retrieval. The characteristics of connectionist networks can be summarized as follows:

1. The structure of connectionist networks resembles that of the human brain much more closely than conventional computers [89];
2. They offer a higher-level configuration in which machine intelligence can follow human intelligence closely in learning, remembering and reasoning [89];
3. These networks are self-processing: once an external force is applied, they process themselves literally. Intelligent behaviours emerge from the local interactions that occur concurrently between the numerous network components [54].

6.1.1 Basic Principles

It is convenient to view any connectionist network as having three constituents: a network, an activation rule, and a learning rule.
The network consists of a set of nodes connected together via directed links. Nodes and links represent different things depending on the motivations for constructing a network. Each node \( i \) in a connectionist network has a numeric *activation level*, \( a_i(t) \), associated with it at time \( t \). Numeric connection strengths, called *weights* and noted as \( w_{ij} \), are often associated with a link between a pair of nodes \( i \) and \( j \) in the network.

The activation rule is a local procedure that each node follows in updating its activation level in the context of input from external sources and neighbour nodes. The total input activation level, \( i_n_i(t) \), that node \( i \) receives from its neighbours and the external sources is used to update its own level of activation, i.e., \( a_i(t+1) = f_i(i_n_i(t), a_i(t)) \); this activation level then communicates from node \( i \) to its neighbours. Thus, a massive, explicit parallelism is involved as activation level spreads throughout a network; all information processing occurs through these local interactions between neighbouring nodes.

Learning in a connectionist network refers to any change that occurs in the behaviour of the network as a result of its experiences over time. Usually, a learning rule is used as a local procedure to describe how the weights on a connectionist network should be altered as a function of time, i.e., \( w_{ij}(t+1) = g_{ij}(a_i(t), a_j(t), w_{ij}(t)) \) [54].

### 6.1.2 History

Work on connectionist networks goes back at least to the 1940s [54]. Early connectionist networks were called neural network models because they tried literally to model networks of brain cells (neurons). Enthusiasm for, and research activity on, the neural modeling waxed during the 1950s and 1960s. However, by the late 1960s, it was decreasing due both to methodological and hardware limitations and to excessively pessimistic attacks on the potential utility of these models. While important work on neural modelling continued during the 1970s, this activity involved a relatively small number of investigators.

Nevertheless, those early research works have laid a very good foundation for further theoretical studies and practices of connectionist networks. Their attempts to study the human brain, their understanding of human intelligent mechanisms, as well as all models built to explain human intelligent behaviours, provide a solid theoretical base which have made connectionist networks applied popularly in other intelligence-related fields.
Another class of connectionist network models emerged in cognitive psychology during the 1960s and 1970s. These early activation spreading models had typically nodes that represented concepts and connections that represented semantically meaningful associations between these concepts. Thus, they are better characterized as associative network models than as neural models. It is not until then that practical applications of connectionist network in other scientific fields have been seen.

### 6.1.3 Connectionist Network & Information Retrieval

Some research has been reported in the literature about the applications of connectionist network in the IR environment, mostly in information retrieval as an alternative to exact-match models such as the Boolean search model (Section 2.2.1). Usually, a three-layer connectionist network is defined for information retrieval: the document layer on the output side, the query layer on the input side, and the index layer as a hidden layer in the middle (Figure 6.1).

![Diagram of a three-layer connectionist network in IR.](image)

**Figure 6.1:** The typical three-layer connectionist network in IR.

In such a connectionist architecture, documents, index terms and queries are all represented as patterns of interconnected nodes whose dynamic structure is manifested in weights associated with the connection links. The theory and practice of the connectionist network representation have often been compared with those of the
conventional vector space model [137]. The research in this area can be analyzed and summarized as follows:

1. Applications of connectionist networks in the IR environment are generally very helpful for the further theoretical studies and practices of IR, because they extend the research of information search and retrieval to another dimension;

2. Applications of the classical three-layer network representation in IR have been fairly well studied. They include Mozer’s first application of a connectionist approach to the areas in IR [54], Cohen & Kjeldsen’s constrained spreading activation in their semantic network for matching research funding agencies and research topics [36], Lelu’s application of spreading activation in image databases [100] and Belew’s AIR project on spreading activation on query-intex-abstract network [8] [131] [132]. These practices are no doubt good experiences for later applications of the network in more complicated IR environments.

3. In theory, because of its inherent associative approach and higher-level structure, activation spreading is closer to human intelligent behaviour and is believed to have the potential to outperform exact-match techniques. Furthermore, as the activation level and the weight defined in connectionist networks are popularly related with probability and conditional probability, the IR models based on connectionist networks are generally able to deal with uncertainty better than the conventional ones. However, the following problems still exist:

   - Experimental verification and comparative evaluation are still needed to fully justify these connectionist network model against other conventional IR models like the vector space model;
   - The classical three-layer connectionist network has undoubtedly been, and is still, the most popular one used in the IR environment. Nevertheless, it does not necessarily represent the best one and needs to be fully justified. Other representation possibilities need to be explored;
   - At the moment, several activation rules are applied both to control the spreading process and to constrain it so that the networks can eventually become stable. These rules need to be fully evaluated;

4. At present, the biggest problem in this research area lies on the adoption of learning and its applications in IR. Traditional learning rules have been applied to the connectionist networks used in IR; however, they do not necessarily represent the most suitable ones. Belew’s AIR project is, by far, the most
thorough study on the subject. His learning is based on the traditional Hebbian learning, but is adapted to the IR environment; therefore, it constitutes the first attempt to introduce a modified correlational learning rule. However, learning in AIR, as in other projects, does not provide any additional information about the Information Space. There, all weights are initialized at the beginning of an IR session and updated by $a_i(t) \times a_j(t)$ while the session goes on [8]. In this way, all auxiliary information provided in the Information Space is only at the activation level of each node. In other words, activation and weight provide the same set of information; they are redundant to each other.

It is shown next that information retrieval is a special area in which new learning and training mechanisms are needed for effective applications of the connectionist network. Its unique characteristics can be listed as follows:

- There exist no universal right answers in IR — what is suitable for one user might not be good for another; even for one user, what is suitable for him/her in one session might not be good for the next — IR is a fully dynamic process in which learning and training should have new meanings. Simple adoption of user-system interactions as training data is not enough;
- As the essence of the learning is to accumulate and inherit past information and experiences, the information collected on different IR sessions should not be ignored and neglected. Traditionally, a librarian can help the library readers to locate the required information more efficiently. This expertise is based on the librarian's knowledge about the information organization in certain libraries, and about the certain reader group he/she is familiar with. The similar knowledge should be equally useful for an automatic information system. For example, statistical information like what the other users have done, or what most of previous users do, may be of value and help for the later users. This information, which could be easily stored on the weight of the related connection, can provide an additional auxiliary information (apart from the activation level for each node) to an Information Space and provide an extra help for on-line navigation.

5. The theoretical studies and practices of connectionist networks in IR are mostly confined to the following limited and small-scale IR environments:

- Bibliographical search and retrieval;
- Abstract search and retrieval.
6. The connectionist network should be most suitable for application of artificial intelligence (AI) in IR. Unfortunately, not much work has been done on this area. The research reported in literature is limited to the following areas:

- Spreading activation on semantic networks has been applied as reasoning mechanisms [139] [67] [159];
- These activation spreading and learning activities are used, mostly for refining search formulation [8].

7. One of the strong points of the connectionist network is that it resembles the human brain much more closely than conventional computers, and its activation spreading and learning mechanism try to mimic those of human intelligence to provide an associative approach. The idea behind the connectionist network theory is very close to that of hypertext, i.e. to emphasis on associations. It is, in fact, an ideal model for the hypertext environment. Unfortunately, until now, little research has been done on this aspect, except for Frisse's work on a highly-structured hypertext environment [67].

Figure 6.2: A connectionist network defined on the Index Space
In the next section, a connectionist network is introduced as the main structure for intelligent behaviours of the soft-link hypertext model (Figure 6.2). Whilst the network by itself is enough in representing the state-space of the conceptual index used in the model, its activation spreading and learning are adopted to control the dynamic operations of the model. For example, its forward activation spreading is represented as making predictions of the user's information needs (user modelling and reasoning) in the soft-link hypertext environment; its backward learning is represented as tuning and improving the conceptual index structure, and accumulating past experiences. In such a way, all events happening on the conceptual index can be well explained, controlled and applied in information retrieval.

The following presentation is in formal specification of Z notation. Based on a typed set theory, Z notation is a language and a style for expressing formal specifications of computing systems [146]. Its key feature is dependent on what is called schema, which consists of a collection of named objects with a relationship specified by some axioms. Z provides notations for defining schemas and later combining them in various ways, so that large specifications can be built up in stages.

One type is defined in this dissertation: [ CONCEPT ]. CONCEPT represents all the concepts in the universe. As shown in Section 4.2, the collection of concepts in any Information Space is a subset of CONCEPT, i.e., \( \text{Concept} \subseteq \text{CONCEPT} \).

For the sake of convenience, the following font and notational conventions are used:

- \( \forall x \in S: P(x) \) means \( (\forall x)[x \in S \implies P(x)] \);
- \( \exists x \in S: P(x) \) means \( (\exists x)[x \in S \land P(x)] \).

Font conventions include:

<table>
<thead>
<tr>
<th>Font</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONCEPT</td>
<td>Type</td>
</tr>
<tr>
<td>ConceptualRelation</td>
<td>Set</td>
</tr>
<tr>
<td>ActivationLevel</td>
<td>Function</td>
</tr>
</tbody>
</table>

The connectionist network is defined in the next section. Its activation rules and learning rules are given in Sections 6.4 and 6.5. These schemas are finally put together, in Chapter 7, to accommodate the soft-link hypertext model.


6.2 The Connectionist Structure in the Soft-Link Hypertext Model.

A connectionist network is now introduced to represent the conceptual index in the Index Space, as well as all events happening in the soft-link hypertext model.

1. A conceptual index, \( G = (V_c, V_e) \), can be defined as a connectionist network. Here, the nodes of the network represent the concepts \( V_c \) in the index; the connections represents the conceptual relations \( V_e \);

2. An activation level, \( a_i(t) \), is assigned to each concept \( V_{ci} (V_{ci} \in V_c) \). In Section 6.4, such an activation level is interpreted as a probability indicator; therefore, at the end of each propagation phase, the activation level of a concept \( V_{ci} \) represents the prediction of the probability that the concept is relevant to the user's information needs;

3. The activation vector, \( a(t) = \langle a_0(t), a_1(t), ..., a_n(t) \rangle \), represents the prediction of the probability of the user's information needs. It reflects the degree that a system understands its user's interests (user model);

4. Activation spreading in the network represents an automatic reasoning and user modelling procedure in the model:
   - When the user expresses his/her information interests by either clicking a concept, or typing in some keywords, one unit of external activation level is interposed into the related node \( (V_{ci} \in V_c) \) in the network;

---

Figure 6.3: The conceptual index represented as a connectionist network.
This activation level spreads in the connectionist network, first to \( V_{ci} \)’s immediate neighbours, and then to its neighbours’ neighbours... From the user’s viewpoint, this is a hypertext navigation procedure, in which the most suitable potential soft link should beat the others and become outstanding at the end of the spreading phase. From the system’s viewpoint, this is a reasoning and user modeling process, in which the system makes predictions about the user’s information needs and next move. Activation spreading process is controlled and constrained by activation rules;

5. A weight is assigned to each connection in the network. It represents the relative strength of the conceptual relation (soft link), compared with other competitors. It is shown later that the weight of a connection, \( V_{ru} = (V_{ci}, V_{cj}) \), refers to the conditional probability that \( V_{cj} \) is relevant given the fact that \( V_{ci} \) is.

6. Unlike the activation level which is initialized at the beginning of each IR session and updated at the every step of the user’s information-seeking process, the weight does not change during one IR session. As a result, two different sets of information are stored and applicable in such a connectionist network:

- Activation level — information about one user during an IR session;
- Weight — statistical information about a group of users in different IR sessions.

7. Learning consists of changing the weights, as result of either the changes of the Information Space, or self-adjustment of the conceptual index. It represents the network’s ability to evolve by experience. Learning is controlled by learning rules.

Figure 6.4 shows a connectionist network, which is composed of concepts and the connections (conceptual relations, or soft links) between these concepts. When a concept (e.g. Hypertext) is chosen by a user, one unit of external activation level is introduced into the node (Hypertext), and therefore, into the network. The activation level spreads to other concepts via connections, so that those of other nodes are also updated. At the end of each propagation phase, the activation level of each concept represents the prediction of the probability that the concept is judged as relevant to the user’s information need. A weight is assigned to each connection, which represents the relative strength of the connection. Learning is a process in which weights of connections are adjusted according to the system’s experience.
More formally, the connectionist network, as well as the rules to control the activation spreading and learning activities, can be presented in a series of schemas. The notations adopted is in Z formal specification language. The advantages of using Z is that it is concise and clear; the mathematical symbols used in Z are stable and well-understood; finally, the loose structure of software specifications can be defined separately as schemas and later combined together in various ways to form a sophisticated model.

First of all, the structure of the connectionist network is demonstrated in the following schema:

<table>
<thead>
<tr>
<th>IndexSpace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept: ( \mathbb{P} ) CONCEPT</td>
</tr>
<tr>
<td>ConceptualRelation: ( \mathbb{P} ) (CONCEPT ( \times ) CONCEPT)</td>
</tr>
<tr>
<td>ActivationLevel: Concept ( \rightarrow ) ( A_C )</td>
</tr>
<tr>
<td>Weight: ConceptualRelation ( \rightarrow ) ( W_R )</td>
</tr>
<tr>
<td>ConceptVisited: Concept ( \rightarrow ) ( N_C )</td>
</tr>
<tr>
<td>ConceptualRelationVisited: ConceptualRelation ( \rightarrow ) ( N_R )</td>
</tr>
<tr>
<td>( A_C ): ( (0, 1) )</td>
</tr>
<tr>
<td>( W_R ): ( (0, 1) )</td>
</tr>
<tr>
<td>( N_C ): ( \mathbb{F} ) ( \mathbb{N} )</td>
</tr>
<tr>
<td>( N_R ): ( \mathbb{F} ) ( \mathbb{N} )</td>
</tr>
</tbody>
</table>

\[
\text{dom ActivationLevel} = \text{dom ConceptVisited} = \# \text{ Concept} \\
\text{dom Weight} = \text{dom ConceptualRelationVisited} = \# \text{ ConceptualRelation} \\
\forall w_{ij} \in W_R: w_{ij} = \frac{1+n_{ij}}{d_C(i)+n_i}
\]
The schema can be explained informally as follows:

- **Concept** is the set of concepts included in the Information Space; it is a subset of all the concepts in the universe, i.e. $\text{Concept} \subseteq \text{CONCEPT}$;

- **ConceptualRelation** is the set of conceptual relations in the Information Space;

- **Concept** and **ConceptualRelation** together accommodate the state-space of the conceptual index in the soft-link hypertext model. In Chapter 4, **Concept** is noted as $V_C$, and **ConceptualRelation** $V_R$ in the conceptual graph $G = (V_C, V_R)$;

- Moreover, all events which happen in the Information Space can be described by the state-space of the network and another four one-to-one functions: ActivationLevel, Weight, Concept Visited and ConceptualRelation Visited;

- ActivationLevel records, in $A_C$, how relevant a concept ($V_{Ci} \in V_C$) is to a user's information needs:
  * $A_C = \{a_i : (0, 1) | 0 \leq i \leq \#\text{Concept}\}$;
  * $a_i = 0$ means that $V_{Ci}$ is totally irrelevant to the user's information needs;
  * $a_i = 1$ means that $V_{Ci}$ is exactly what the user is looking for.

- Weight records, in $W_R$, the relative importance of a conceptual relation ($V_{Rij} \in V_R$) for the information user:
  * $W_R = \{w_{ij} : (0, 1) | 0 \leq i, j \leq \#\text{Concept}\}$;
  * $w_{ij} = 0$ means that there is no connection between $V_{Ci}$ and $V_{Cj}$;
  * $w_{ij} = 1$ means that $V_{Rij}$ is the only connection between $V_{Ci}$ and $V_{Cj}$.

- Concept Visited records, in $N_C$, how many times a concept has been visited.
  * $N_C = \{n_i : n_i \geq 0 | 0 \leq i \leq \#\text{Concept}\}$

- ConceptualRelation Visited records, $N_R$, how many times a conceptual relation ($V_{Rij} \in V_R$) has been visited.
  * $N_R = \{n_{ij} : n_{ij} \geq 0 | 0 \leq i, j \leq \#\text{Concept}\}$

- For each conceptual relation $V_{Rij} = (V_{Ci}, V_{Cj})$, the relationship between $w_{ij}$, $n_i$ and $n_{ij}$ can be expressed as:
  * $w_{ij} = \frac{1 + n_{ij}}{d^+_C(i) + n_i}$, where $d^+_C(i)$ is the out degree of $V_{Ci}$ in graph $G$.

- The predicate part of the schema implies:
  * $\sum_j w_{ij} = 1$, where $i = 1, 2, \ldots, \#\text{Concept}$. 
Before the connectionist network is further developed and applied in the soft-link hypertext model, it is necessary to discuss the four important variables defined in the network. They are the weight \( w_{ij} \), concept Visited \( (n_i) \), ConceptualRelationVisited \( (n_{ij}) \) and activation level \( a_i \). Such an investigation should help to explain the motivation for introducing the connectionist network into the dissertation.

First of all, it is assumed that every time a user invokes the hypertext mechanism, one and only one of the hypertext associations (soft link) provided is to be followed.

**Weight\(^1\)**

As suggested in Figure 6.4, every directed connection from a concept \( V_{Ci} \) to one of its \( d_C^j(i) \) immediate successors represents an association, or a soft hypertext link. Every time a user invokes the soft-link hypertext mechanism by clicking \( V_{Ci} \), he/she needs to choose one of these links to follow. Therefore, the question becomes, among these \( n \) soft links \( [n = d_C^j(i)] \), which one should be most relevant to the user’s information needs, or by following which link the user would most likely locate the information required. This is the similar issue to that of the probabilistic model (Section 2.2.1).

As in the probabilistic model \([130]\), there exists no mechanism which tells us without fail which hypertext link is relevant and which is not. In a hypertext environment, the only way relevance can ultimately be decided is for the user to follow a hypertext link, retrieve the destination component and read the full text. Nevertheless, as in the conventional probabilistic information retrieval environment, imperfect knowledge can be used to guess for any given hypertext link, \( V_{Rij} \), whether it is relevant or not for a hypertext request \( V_{Ris} \). Here the weight associated with each connection in the network collects the statistical information; it is used to to provide sensible evidence to estimate, for any request \( V_{Ris} \), the probability that it is identical to a conceptual connection \( V_{Rij} \) in the network:

\[
P(V_{Ris} = V_{Rij} | V_{Ris})
\]

Starting at any concept \( V_{Ci} \) as a hypertext source anchor, there exist \( n \) soft hypertext links to choose and follow: \( V_{Ri1}, V_{Ri2}, ... V_{Rin} \), where \( n = d_C^j(i) \). As the result of

\(^1\)The discussion of probability and related notations in this section follow those in Rijssbergen's book, Information Retrieval, pp. 112 – 115 [126].
the assumption, for any soft-link request $V_{Riz} = (V_{Ci}, V_{Cz})$, there are $n$ mutually exclusive events.

$E_1 = \text{The requested link } V_{Riz} \text{ is } V_{Ri1};$

$E_2 = \text{The requested link } V_{Riz} \text{ is } V_{Ri2};$

...  

$E_n = \text{The requested link } V_{Riz} \text{ is } V_{Rin}.$

With Bayes' Theorem, the probability, $P(V_{Riz} = V_{Rij} \mid V_{Riz})$ is calculated as:

$$P(E_j|V_{Riz}) = \frac{P(V_{Riz}|E_j)P(E_j)}{P(V_{Riz})}$$ (2)

Here, $j = 1, 2, \ldots, n$;

$P(E_j)$ is the prior probability of the event $E_j$;

$P(V_{Riz} \mid E_j)$ is proportional to what is commonly known as the likelihood of an event $E_j$ given $V_{Riz}$;

$P(V_{Riz})$ is the probability of observing $V_{Riz}$ on a random basis given that it may be any of these events:

$$P(V_{Riz}) = \sum_{j=1}^{n} P(V_{Riz}|E_j)P(E_j)$$ (3)

Finally, the following equation always exist:

$$P(E_1|V_{Riz}) + P(E_2|V_{Riz}) + \ldots + P(E_n|V_{Riz}) = 1$$ (4)

Equation (2) shows a calculation, or estimation, of the probability that for any soft-link request $V_{Riz}$, $V_{Riz}$ is $V_{Rij}$. The calculation is based on Bayes' Theorem, which relates the posterior probability of relevance to the prior probability of relevance and the likelihood of relevance after observing a random soft link request. Next, it is necessary to examine the weight defined in the connectionist network and find out what it means. For any conceptual relation $V_{Rij} = (V_{Ci}, V_{Cj})$, the relationship between its weight ($w_{ij}$), the times $V_{Ci}$ has been visited as the source anchor ($n_i$) and the times $V_{Rij}$ has been used ($n_{ij}$), has been expressed as:
\[ w_{ij} = \frac{1 + n_{ij}}{d_C^+(i) + n_i} = \frac{1 + n_{ij}}{d_C^+(i) + \sum_{k=1}^n n_{ik}} = \frac{1 + n_{ij}}{d_C^+(i) + n_{i1} + n_{i2} + \ldots + n_{in}} \]

Comparison between Equations (2) and (5) leads immediately to the following conclusion: \( w_{ij} \) represents an attempt to estimate the probability that a soft-link request \( V_{R_i} \) is \( V_{R_j} \). This estimation is based on the statistical information about the relative popularity of \( V_{R_j} \) (\( n_{ij} \) vs. \( n_i \)).

1. When \( n_i = 0 \), all the connections have the same probability, i.e. \( w_{ij} = \frac{1}{d_C^+(i)} \).

2. When \( n_i \rightarrow \infty \), \( w_{ij} \) should approximate \( P(E_j|V_{R_i}) \).

\( n_i \) and \( n_{ij} \)

In the connectionist network, \( n_i \) records how many times a concept has been visited. It, therefore, provides the statistical information about the total number of samples for the observed group of links. This information represents the probability of observing any hypertext request \( V_{R_i} \) on a random basis, given that it may be any of the \( n \) events \( (E_1, E_2, \ldots, E_n) \), as expressed in Equation (3).

\( n_{ij} \) records the times a soft hypertext link, \( V_{R_{ij}} \), has been visited. It provides the statistical information that the event \( E_j \) is true amongst the total samples \( n_i \).

**Activation Level**

For each concept \( V_{Gi} \) in the network, there exists a numerical measurement, called the activation level \( (a_i) \), attached to it. \( a_i \) is used to reflect the importance of the concept \( V_{Gi} \) for the present user’s information needs. It is set to zero at the beginning of the session and updated whenever the user interacts with system. It is shown in Section 6.4 that, in fact, \( a_i \) reflects the probability that a concept is relevant to the user's information needs. The activation level is used in the soft-link hypertext model for on-line navigation and reasoning.
In the next three sections, three groups of rules are developed to control the dynamic operations of the connectionist network. These rules are first presented in a series of schemas. They are applied in Chapter 7 to facilitate intelligent mechanisms in the soft-link hypertext model. These three groups of rules include the following:

1. Conceptual index formulation;
   This defines the rules for the formulation of the conceptual index. It includes the automatic generation of the global index and the auxiliary semantic net for the conceptual index, and updating the data structure when the Information Space changes. Whilst the related mechanism has been discussed in Chapter 5, its functions are summarized more concisely and clearly in Section 6.3.

2. Forward activation spreading;
   Based on the connectionist network defined, its forward activation spreading is applied in the soft-link hypertext model to reason and navigate during information retrieval. Activation rules used to constrain such a spreading process are introduced as two schemas in Section 6.4.

   The backward learning in the connectionist network is used to collect the statistical information of the application of the conceptual index, so that it can be used to self-adjust the data structure, and provide an auxiliary information for the later users. The learning rules used are also represented as two schemas in Section 6.5.

6.3 Automatic Index Formulation Mechanism

Maintenance of the soft-link hypertext model includes formulating the conceptual index at initialization, and updating the data structure when the Information Space changes. It is facilitated by an automatic index formulation mechanism used in the model (Chapter 5). The related events include the following three categories of activities, which are presented in five schemas:

1. Index initialization (1 schema);
2. Adding concept(s) and conceptual relation(s) in the Information Space (2);
3. Deleting concept(s) and conceptual relation(s) in the Information Space (2).
6.3.1 Initialization — Before Indexing

Before the index formulation starts, the Concept in the Index Space is empty.

\[
\begin{array}{l}
\text{InitIndexSpace} \\
\Delta \text{IndexSpace} \\
\text{Concept} = \phi
\end{array}
\]

6.3.2 Indexing — Addition of Concepts and Conceptual Relations

As shown in Chapter 5, formulation of the conceptual index refers to identification a global index structure and the related semantic net. The former facilitates the global indexing process (identification of mapping from keywords to documents), which has been well studied; the latter represents the hyperization process (identification of associations among concepts). In the soft-link hypertext model, this hyperization process is changed into the formulation of a special semantic net. Expressed in the formal specification, these two steps are represented by two schemas: AddConcept and AddConceptualRelation.

\[
\begin{array}{l}
\text{AddConcept} \\
\Delta \text{IndexSpace} \\
V_{Ci}? \colon \text{CONCEPT}
\end{array}
\]

\[
\begin{array}{l}
V_{Ci} \notin V_C \\
\text{Concept}' = \text{Concept} \cup \{V_{Ci}\} \\
d_{Ci}'(i) = 0 \\
\text{ActivationLevel}' = \text{ActivationLevel} \cup \{V_{Ci} \mapsto 0\} \\
\text{ConceptVisited}' = \text{ConceptVisited} \cup \{V_{Ci} \mapsto 0\}
\end{array}
\]

Schema AddConcept defines the event that a new concept is added into the index:

- When a new concept \( V_{Ci}? \) is introduced, it is included in \( \text{Concept} \);
- Its activation level is defined as zero, i.e., \( a_i = 0 \);
- Upon introduction, the time \( V_{Ci}? \) has been visited is equal to zero, i.e., \( n_i = 0 \).
AddConceptualRelation

\[ \Delta \text{IndexSpace} \]
\[ V_{Rij}? : (\text{CONCEPT} \times \text{CONCEPT}) \]

\[ V_{Ci} \in V_C \land V_{Cj} \in V_C \land V_{Rij}? \notin V_R \]
ConceptualRelation ' = ConceptualRelation \cup \{V_{Rij}?\}
ConceptualRelationVisited ' = ConceptualRelationVisited \cup \{V_{Rij}? \mapsto 0\}
\[ d_C^+ (i)' = d_C^+ (i) + 1 \]
\[ \forall w_{ik} \in W_R : w_{ik}' = \frac{1 + n_{ik}}{d_G^+(i)' + n_i}, \quad (k = 0, 1, ..., d_C^+(i)' and k \neq j) \]
Weight ' = Weight \cup \{V_{Rij}? \mapsto \frac{1}{d_G^+(i)' + n_i} \}

Schema \textit{AddConceptualRelation} defines the event that a new conceptual relation is introduced into the conceptual index:

- No conceptual relation can be added into the Index Space unless both concepts connected by the conceptual relation are in \( V_C \). [This is not a problem for the soft-link hypertext model, as the global indexing, which facilitates the addition of concept (AddConcept), always happens before the formulation of the semantic net (AddConceptualRelation) starts];
- When a new conceptual relation \( V_{Rij}? \) is first built up, \( n_{ij} = 0 \).

6.3.3 Removal — Deletion of Concepts and Conceptual Relations

As opposite to the index formulation procedure, no concepts can be removed from the conceptual index until all conceptual relations between it and its neighbours are removed first. The operations in this category include \textit{RemoveConceptualRelation} and \textit{RemoveConcept}.

RemoveConceptualRelation

\[ \Delta \text{IndexSpace} \]
\[ V_{Rij}? : \text{ConceptualRelation} \]

\[ \text{ConceptualRelation}' = \text{ConceptualRelation} \setminus \{V_{Rij}?\} \]
\[ \text{ConceptualRelationVisited}' = \text{ConceptualRelationVisited} \setminus \{V_{Rij}? \mapsto n_{ij}\} \]
\[ \text{Weight}' = \text{Weight} \setminus \{V_{Rij}? \mapsto w_{ij}\} \]
\[ d_C^+ (i)' = d_C^+ (i) - 1 \]
\[ n_i' = n_i - n_{ij} \]
\[ \forall w_{ik} \in W_R : w_{ik}' = \frac{1 + n_{ik}}{d_G^+(i)' + n_i'}, \quad (k = 0, 1, ..., d_C^+(i)' and k \neq j) \]
• When a conceptual relation $V_{Rij}$ is removed from the Index Space, it is deleted from the ConceptualRelation. The related $n_{ij}$ and $w_{ij}$ are removed, and the related $n_i$ is updated;

• The weights of other related connections are adjusted accordingly.

<table>
<thead>
<tr>
<th>RemoveConcept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ IndexSpace</td>
</tr>
<tr>
<td>$V_{Ci}$?: CONCEPT</td>
</tr>
</tbody>
</table>

$V_{Ci}? \in V_G \land \exists V_{Cij} \in V_G : (V_{Rij} \in V_R \lor V_{Rji} \in V_R)$

$Concept' = Concept - \{V_{Ci}\}$

$ActivationLevel' = ActivationLevel - \{V_{Ci} \mapsto a_i\}$

$ConceptVisited' = ConceptVisited - \{V_{Ci} \mapsto n_i\}$

• For the schema RemoveConcept, the predicate part first checks if there are any conceptual relations ending on $V_{Ci}?$;

• If not, the concept is removed from the Concept, so are $a_i$ and $n_i$.

### 6.4 Activation Level and Activation Rules.

For each concept $V_{Ci}$ in the network, there is a numerical measurement (the activation level, or $a_i$) attached to it. The application of the conceptual index in IR is represented by the continuous updating and spreading of the activation levels in the network.

At the beginning of an IR session, the activation level of each concept in the network is set to zero, i.e. $\forall a_i \in A_C : a_i = 0$. Then, the session starts: Each time the user interacts with the system by typing in a word or clicking an anchor, one unit of external activation is inputed into the related concept. As a result, this external activation input stimulates the network; one unit of activation level spreads from the related concept to its immediate successors, and then to its successors' successors...

Thus, the activation level of all concepts in the network is updated. At the end of each spreading phase, an updated sequence is outputted, It offers a list of concepts in the decreasing order of their relevancy ($a_i$). This process repeats till the end of the session.
Activation spreading in the soft-link hypertext model is controlled by the following four activation rules:

1. External Activation Rule. This is to control the external stimulation process. For any concept $V_i$, the external input activation level it receives is noted as $i_{n_i}^+(t)$.

   
   
   
   
   $i_{n_i}^+(t) = \begin{cases} 1 & \text{if the concept } V_i \text{ is chosen, at time } t, \text{ as an external request;} \\ 0 & \text{otherwise} \end{cases}$

   
   
   
   

   Here, the external inputs may come from the following outside stimulation sources:

   
   
   
   
   • The Information Interests explicitly expressed;

   
   
   
   
   • The keywords used in queries;

   
   
   
   
   • The keywords clicked as hypertext anchors.

2. Spreading Activation Rule. A concept's activation level, or output derived from it, is communicated from the concept to its successors (both conceptual and semantic), and to the successors' successors... This process is called activation spreading. In the soft-link hypertext model, the spreading activation rule is what each concept follows in updating its activation level in the context of input from its immediate predecessor concepts, as the result of the activation spreading procedure. The spreading activation rule is expressed as:

   
   
   
   
   $i_{n_i}^-(t) = \sum_j a_j(t) \times w_{ji}(t), \quad \text{where } i = 0, 1, \ldots, n.$

3. Input Activation Rule. The total input activation level for a concept $V_i$ includes the external input activation level $i_{n_i}^+(t)$ and spreading activation level $i_{n_i}^-(t)$, i.e.,

   
   
   
   
   $i_{n_i}(t) = i_{n_i}^+(t) + i_{n_i}^-(t), \quad \text{where } i = 0, 1, \ldots, n.$

4. Output Activation Rule. The final activation level of a concept $V_i$, i.e., the output from it, is updated by the following rule.

   
   
   
   
   $a_i(t) = a_i(t - 1) + i_{n_i}(t), \quad \text{where } i = 0, 1, \ldots, n.$

The applications of the soft-link hypertext model in IR can be described by the following two schemas in Z notation:
Session Starts

Δ IndexSpace
PATH*: \( \mathbb{F} \) CONCEPT

\[ \forall a_i \in A_C: a_i = 0 \]
\[ PATH^* = \phi \]

Forward Spreading

Δ IndexSpace
PutInOrder: \((Concept \times \text{decoder}[A_C]) \rightarrow \text{seq Concept}\)
PATH*: \( \mathbb{F} \) CONCEPT

\[ V_{Gi}? : \text{Concept} \]
rep!: seq Concept

\[ V_{Gi}? \in V_C: a_i' = a_i + 1 \]
\[ \forall V_{Ci} (V_{Ci} \in V_C \land V_{Ci} \notin PATH^*): a_n' = a_n + \sum_m w_{mn} \times a_m \]
\[ \forall V_{Ci} \in PATH^*: a_n' = 1 \]
\[ PATH'^* = PATH^* + \{V_{Ci}?\} \]

Here, \( PATH^* \) is the collection of the user's browsing trail, i.e. History (Chapter 4). The function PutInOrder takes the set Concept and converts it into a sequence, rep!. rep! contains all the elements of Concept, ordered according to the decreasing order of \( a_i \). At each step during an IR session, the sequence rep! is used as the main interactive mechanism between the user and the soft-link hypertext system. Concept included in the Information Space are divided into the following two groups:

- **Group 1** — \( \forall V_{Gi} \in PATH^*: a_i = 1 \).
  This group includes all the concepts (nodes) which have been visited by the user during the session; these concepts reside on the top of rep!. As a result, it is always easy for the user to review any of the visited concepts and go back to any of these concepts;

- **Group 2** — \( \forall V_{Ci} (V_{Ci} \in V_C \land V_{Ci} \notin PATH^*): a_i < 1 \).
  In rep!, concepts in Group 2 come after those in Group 1. For these concepts, rep! provides an ordered list according to the prediction that a concept may be relevant to the user's information interests. This sequence is presented to the user, in order to speed up the information search process. The higher a concept
6.5 Improving the Conceptual Index Structure

Learning (adaptation) refers to any change that occurs in the behaviour of the model as a result of its experiences over time; in the connectionist network, a set of learning rules is needed to control such a learning procedure [54]. In the soft-link hypertext model, this change of behaviour is represented by long-term memory (LTM) and short-term memory (STM). Whereas LTM refers to the changes that occur in the behaviour of the model as a result of its experiences over different IR sessions by different users, STM records one specific user's information needs and browsing trail during one IR session. As a part of the connectionist network, the learning mechanism and its learning rules concerning with LTM is discussed in the section, whereas STM and its relation with LTM are covered in detail in Section 7.4.

Here, the backward learning in the connectionist network is used in the soft-link hypertext model to improve the quality of the conceptual index via statistical information. As a result, the soft-link hypertext model can self-organize its data structure, and evolve in such a way that the weights of these conceptual relations are slowly tuned to represent the general popularities of the related soft links. This procedure is controlled by backward learning on the connectionist network, i.e., whenever a concept \( V_{Ci} \in V_C \) and conceptual relation \( V_{Rij} \in V_R \) is used, \( n_i \) and \( n_{ij} \) are updated so that all weights of the conceptual relations related with \( V_{Ci} \) are changed. Noted in Z specification, there are \textit{BackwardLearning-1} and \textit{BackwardLearning-2}:

\[
\begin{array}{|l|}
\hline
\text{BackwardLearning-1} \\
\Delta \text{IndexSpace} \\
V_{Ci}? : \text{IF CONCEPT} \\
\hline
V_{Ci}? \in V_C \land a_i = 1 \\
\quad \land \exists V_{Rij} \in V_R: \\
n_i' = n_i + 1 \\
\hline
\end{array}
\]

Schema \textit{BackwardLearning-1} defines the learning procedure after a concept is chosen as a hypertext source anchor. If there exists a hypertext soft link, then \( n_i' = n_i + 1 \).
BackwardLearning-2

\[ \Delta \text{IndexSpace} \]

\[ V_{Cj}? : \text{F CONCEPT} \]

\[ V_{Cj}? \in V_C \land a_i = 1 \land a_j = 1 \]
\[ \land \exists V_{Rij} \in V_R: \]
\[ n_{ij}' = n_{ij} + 1 \]
\[ \forall w_{ij} \in W_R: w_{ij}' = \frac{1+n_{ij}'}{d_C'(i)+n_i} \]

Schema BackwardLearning-2 defines the learning procedure when a concept is chosen as a hypertext destination anchor. Here, a soft link \( V_{Rij} \) has been used by the user so that its relative importance, compared with other related links, should be updated.

6.6 Further Discussions on the Learning Rule

In the soft-link hypertext model, the weight related with each soft link is constantly updated, according to the relative times the soft link has been used:

\[ w_{ij} = \frac{1+n_{ij}}{d_C'(i)+n_i} \]  \hspace{1cm} (6)

The idea behind such a learning process is that the more frequently a soft link is chosen by the user, the more important such an association should be for the user population; thus, its weight should become heavier to reflect the popularity. It is shown in this section that such a learning process constitutes a special group of competitive (Kohonen) learning. It can also be very useful in other applications.

6.6.1 Kohonen Layer and Competitive Learning

In neural networks, learning rules belonging to the competitive learning (also called Kohonen learning) category all have the property that a competition process, involving some or all of the processing elements of the neural network, always takes place before each episode of learning. The processing elements that emerge as the winners of the competition are then allowed to modify their weights (or modify their weights in a different way from those of the non-winning units).
A Kohonen layer is shown in Figure 6.5. The basic idea of Kohonen learning is to have a layer of processing elements arrange their weight vector such that these weight vectors are distributed in $\mathbb{R}^n$ with a number density approximately proportional to the probability density function $\rho$ according to which the $x$ input vectors used to train the layer are selected. Therefore, it constitutes a self-organizing and training procedure.

The $n$ Kohonen processing elements of this layer each receives $n$ inputs $z_1, z_2, \ldots, z_n$. Each input has a weight $w_{ij}$ assigned to it. When each $x$ vector is entered into the Kohonen layer, the processing elements compete on the basis of which of them has its weight vector $w_i$ [$w_i = (w_{i1}, w_{i2}, \ldots, w_{in})$] closest to $x$ (as measured by a distance function $D$). The winner emits signal $z_i = 1$; the others emit $z_i = 0$. Furthermore:

1. The training data for the Kohonen learning is assumed to consist of a sequence of input vectors $x$, drawn at random in accordance with a fixed probability density function $\rho$;

2. As each of these vectors is entered into the network, the Kohonen processing elements compete with one another to determine the winner on the basis of minimum distance:

$$|x - w_c| = \min_i |x - w_i|$$

Figure 6.5: A Kohonen layer.
3. The winning Kohonen processing element then has its \( z_c \) set to 1, and all of the other Kohonen unit output signals are set to 0;

4. Weight modification takes place in accordance with the following equation, called the **Kohonen learning law**:

\[
 w_i^{\text{new}} = w_i^{\text{old}} + \alpha(x - w_i^{\text{old}})z_i 
\]  

- For the winning connection, \( z_c = 1 \). Therefore, this learning rule can be rewritten as:

\[
 w_i^{\text{new}} = (1 - \alpha)w_i^{\text{old}} + \alpha x 
\]  

- For the losing connections, \( z_i = 0 \). Therefore, this learning law will be:

\[
 w_i^{\text{new}} = w_i^{\text{old}} 
\]

- \( \alpha \) is the modification rate. At the beginning of training, it is often set to a value of approximately 0.8. As the \( w_i \) vectors move into the area of the data, \( \alpha \) is then lowered to 0.1 or less for final equilibration [86];

5. As training progresses, the Kohonen weight vectors become densest where the \( x \)s are most common, and become least dense (or absent) where the \( x \) vectors hardly ever (or never) appear. In this way, the Kohonen layer adapts itself to conform approximately to \( \rho \) in a volume number density sense [86];

6. Finally, Kohonen proved the following proposition [96]:

*Starting with randomly chosen initial values for the \( w_i \), these numbers will gradually assume new values in a process specified by Equations (7) - (10), such that as \( t \to \infty \), the set of numbers \( (w_1, w_2, ..., w_n) \) becomes ordered in an ascending or descending sequence. Once the set is ordered, it remains so for all \( t \). Moreover, the point density function of the \( w_i \) will finally approximate \( \rho(x) \).*

### 6.6.2 The Learning Rule Used in the Soft-Link Hypertext Model.

The learning rule used in the soft-link hypertext model shares a lot of similarities with the Kohonen learning rules, although it has its own unique characteristics. In fact, each group of the local associations in the semantic net constitutes a one-dimensional Kohonen layer. If Figure 6.3 (which is part of Figure 6.4) is reorganized, it can be
represented in Figure 6.6. It can be seen that, in a conceptual index, the connections between each concept and its immediate successors constitute a Kohonen layer. Here, only one input $x$ is put into the layer, which represents the conceptual path (a soft hypertext association) the user has just visited. The related conceptual connection becomes the winner of the competition. This can be further explained as follows:

1. For each $V_{Ci} \in V_C$, there exists a Kohonen layer, which consists of $n$ processing elements $[n = d_C(i)]$. Each of these elements is a conceptual connection, $V_{Rij}$, from $V_{Ci}$ to and one of its immediate successors, $V_{Cj}$. Each processing element (conceptual connection) has a weight $w_{ij}$ associated with it;

2. At any moment, there is only one input $x$ in the layer. It represents the conceptual connection (soft link) being visited;

3. The training data $X$ for the learning process consist of a sequence of inputs $x$, each of which represents a conceptual connection (soft hypertext link) being used. As these inputs record each individual user's browsing in the Information Space, they are, in fact, random samples in natural accordance with a fixed probability density function $\rho$. This density function reflects the general popularity of the related hypertext links for the user population. Although the probability density function $\rho$ is unknown to anybody, this training procedure is able to self-organize the layer to represent it — tuning the weights of conceptual connections (soft hypertext links) so that they could eventually reflect the
general popularity of these related links, this is exactly the purpose of learning in the soft-link hypertext model.

4. At any moment, there is one winner $w_{ik}$ in the layer,

$$|x - w_{ik}| = 0 \quad (11)$$

5. The winning processing element adapts its weight as the following:

$$w_{ik}^{new} = \frac{1 + n_{ik}^{new}}{d_C^+(i) + n_{ik}^{new}}$$

$$= \frac{1 + (n_{ik}^{old} + 1)}{d_C^+(i) + (n_{ik}^{old} + 1)}$$

$$= \frac{2 + n_{ik}^{old}}{d_C^+(i) + n_{ik}^{old} + 1}$$

$$= \frac{1 + n_{ik}^{old}}{d_C^+(i) + n_{ik}^{old}} + \frac{d_C^+(i) + n_{ik}^{old} - n_{ik}^{old} - 1}{(d_C^+(i) + n_{ik}^{old})(d_C^+(i) + n_{ik}^{old} + 1)}$$

$$= w_{ik}^{old} + \frac{d_C^+(i) + n_{ik}^{old} - n_{ik}^{old} - 1}{(d_C^+(i) + n_{ik}^{old})(d_C^+(i) + n_{ik}^{old} + 1)} \quad (12)$$

Equation (12) is very similar to Equation (9). The only difference is that, instead of moving $w_{ik}^{new}$ a small portion towards $x$ (the training data) as shown in Equation (9), its value is increased a small portion to enhance this winning position. With Equation (11), it is impossible to adjust locations of the weight in the vector space to approach the training data, as in the conventional Kohonen learning. Instead, the value attached to each weight is adjusted here to represent this density function so as to approach the training data.

6. Other conceptual connections adapt their weights as the following:

$$w_{ij}^{new} = \frac{1 + n_{ij}^{new}}{d_C^+(i) + n_{ij}^{new}}$$

$$= \frac{1 + n_{ij}^{old}}{d_C^+(i) + (n_{ij}^{old} + 1)}$$

$$= \frac{1 + n_{ij}^{old}}{d_C^+(i) + n_{ij}^{old} + 1}$$

$$= \frac{1 + n_{ij}^{old}}{d_C^+(i) + n_{ij}^{old}} - \frac{n_{ij}^{old} + 1}{(d_C^+(i) + n_{ij}^{old})(d_C^+(i) + n_{ij}^{old} + 1)}$$

$$= w_{ij}^{old} - \frac{n_{ij}^{old} + 1}{(d_C^+(i) + n_{ij}^{old})(d_C^+(i) + n_{ij}^{old} + 1)} \quad (13)$$
7. All weights in the layer are modified in such a way that the weight of the winning connection increases a bit, whereas those of the losing ones decrease a bit. The sum of all weights in the layer stays constant (equal to 1) at any moment, which is in accordance with the learning rules (Section 6.2).

\[ \sum_j \Delta w_{ij} = 0 \]  

8. The modification rate, \( \alpha \), decreases as the training progresses:

\[ \alpha \propto \frac{1}{(d_G^+(i) + n_t^{old})(d_G^-(i) + n_t^{old} + 1)} \]  

9. It can be seen that the learning rule used in the soft-link hypertext model is a standard one-dimensional Kohonen learning rule. With the proven proposition for the Kohonen layer, it can, therefore, be concluded that starting with the initial values for \( w_{ij} \), (which equal to \( \frac{1}{d_G^+(i)} \) at initialization), these weights will gradually assume new values in a process specified by the learning rule (Section 6.5), such that as \( t \to \infty \), the set of numbers becomes ordered. Once the set is ordered, it remains so for all \( t \). Moreover, \( w_{ij} \) will finally approximate \( \rho(x) \).

6.6.3 The Unique Characteristic and Its Generalization

Although, as discussed above, learning in the soft-link hypertext model is a standard Kohonen self-organization process, it has its unique characteristics:

1. It is different from other reported Kohonen learning in that its training data are not drawn from a well-defined domain;

   Traditionally, these training data are drawn at random from a known domain at a fixed probability density function \( \rho \), and fed into the Kohonen layer to train the system [86]. However, in the soft-link hypertext model, although it is assumed that there exists a user population and exists some kind of probability density function \( \rho \) reflecting the user's preference of these hypertext links, this \( \rho \) function is unknown. In other words, our understanding of the target domain in IR (the user's preference of hypertext links) is not as good as in other traditional Kohonen learning process.
2. Unlike the traditional Kohonen learning, sampling and training in IR is uncontrollable;

3. Instead of self-organizing the relative position of weight vectors in the traditional Kohonen layer [moving the weight vector a fraction $\alpha$ of the way along the straight line from the old weight vector to the $x$ vector as shown in Equation (9)], the learning rule used in the soft-link hypertext model adjusts the value of weights to approach the probability density function $\rho(x)$. Here, the training data $xs$ are hypertext links themselves. Therefore, Equation (11) always exists. This characteristic enables the learning rule used in the soft-link hypertext model to be effective even in the extremely unevenly distributed probability density function situation;

4. All efforts to avoid the over-learning (e.g., radial sprouting, noise vectors, and conscience of processing elements [89]) are unnecessary here.

6.7 Summary

In this chapter, a connectionist network has been introduced as a formal theoretical basis for presentations of intelligent mechanisms in the soft-link hypertext model. Based on the network, it has been shown how the forward activation spreading can be used to make predictions of the user's information needs in the soft-link hypertext environment. Such a predicting process is further developed, in Chapter 7, into a user modelling and reasoning mechanism for the soft-link hypertext model. It has also been shown how the backward learning of the connectionist network can be applied to tune and improve the conceptual index structure formulated in Chapter 5. This learning mechanism represents a slow training and self-adjusting procedure through which the weights associated with connections could eventually be tuned to reflect the general popularity of the related soft links. This could provide some valuable information to help the user during information search and retrieval. Furthermore, the learning rules adopted in the model constitute a special group of competition learning (Kohonen learning), which can be extended into more general applications. The connection network defined in this chapter is used, in Chapter 7, to facilitate intelligent mechanisms in the soft-link hypertext model. This model is implemented and evaluated in Chapters 8 and 9.
Chapter 7
Intelligent Mechanisms in the Soft-Link Hypertext Model

This chapter first presents uncertainty factors in IR. The intelligent mechanisms used in the soft-link hypertext model are then developed. These mechanisms are designed to deal with uncertainty in the hypertext environment. There are four different, but interrelated mechanisms: index formulation (which has been discussed in Chapter 5), learning, user modelling and reasoning. The soft-link hypertext model supports an implicit, continuous and dynamic user modelling. Its learning mechanism includes both long-term memory (LTM) and short-term memory (STM). Finally, the reasoning mechanism makes prediction, based on its user model and its experiences (LTM and STM), of probability that each concept is relevant to the user's information needs. Whilst the theoretical basis for these mechanisms has been developed in Chapter 6, it is applied in this chapter to facilitate the intelligent soft-link hypertext model.

7.1 Uncertainty

An essential aspect of any intelligent system is the need to deal with uncertainty. Uncertainty can arise because of inherent imprecision in the rules used in those systems, or because of imprecision in the data fed into the systems. Inclusion of facilities for handling imprecise rules or data is a defining feature of intelligent systems.

Various theories have been developed in IR to accommodate uncertainty. Certainty factors, confidence factors, or probabilities associated with the information components indicate the extent to which a system believes the information components are relevant to the information user's needs. They are used during reasoning processes to express a degree of confidence in the conclusions reached [115] [20].

In this chapter, uncertainty in various information retrieval models is firstly analyzed. After that, four intelligent mechanisms, based on the connectionist network defined in the Index Space (Chapter 6), are developed. As the result, a Document Space, a well-structured conceptual index and four auxiliary intelligent mechanisms constitute together the soft-link hypertext model.
7.2 Uncertainty in IR

In IR, uncertainty exists when the SIMILAR mechanism searches for relevant documents in the Document Space to meet the user's information needs. Uncertainty is mainly caused for the following reasons:

1. Search formulation: imprecise input data are used to represent the information needs, for one or more of the following reasons:
   - The information users often do not know exactly what they want;
   - Though the users may know what they want, they are often not able to present the information needs properly;
   - Though the information needs could be properly presented, they are not understandable (or misunderstood) by an IR system.

2. Indexing: improper identifiers are adopted to represent the Document Space;

3. Searching: unsuitable searching mechanisms are used in a system.

7.2.1 Traditional Search and Retrieval

In the traditional IR environment, uncertainty exists in all three aspects of information retrieval, including indexing, search formulation and searching. The following strategies are used in traditional IR systems to deal with uncertainty, to reason in uncertain circumstances and to improve the quality of information retrieval:

1. Usually, the prediction of probability, instead of crisp yes or no, that a document is judged as relevant to the information user's needs is assigned and attached to each document retrieved [99]. This prediction (often in the form of weight) provides a belief indicator for the related retrieval set in the uncertain environment. Different weighting algorithms are applicable [134];

2. Confirmation of certainty factors can be used, as feedbacks or dynamic query adjustments, for search reformulation or retrieval refinements [134];

3. Various semantic networks, including the Bayesian network and belief networks [139] [67] [169], have been used to adjust search formulations (thesaurus) and to reason during searching process.
7.2.2 Traditional Hard-Link Hypertext

Traditional hard-link hypertext is a special case for uncertainty. The reasons are listed as follows:

1. As hypertext users express their information needs by clicking hypertext source anchor, or choosing hypertext options, uncertainty caused by search formulation is effectively avoided in hypertext. In fact, utilization of the recognition, rather than the recall, is one of the main advantages of hypertext;

2. However, the implicit link structure adopted in the hard-link hypertext model and the conventional manual hyperization process used to generate such a link structure brings uncertainty factors. The main reasons are two-fold:
   • The pre-set hypertext structure usually represents only the authors’ subjective, personal, or even biased views in expressing the interrelationships among information components. It may not necessarily be the best;
   • Authors’ and editors’ motivations for certain settings, as well as the link structure itself, are usually not explicit (Chapter 3); they may not always be well understood by the information user.

As the result, hypertext users quite often retrieve some information by following these preset links, only to find that the retrieved component is not what they are seeking.

3. What makes it worse is that, in the traditional hard-link hypertext environment, the searching process becomes browsing in an Information Space, and the SIMILAR mechanism is totally left to the users; thus, dealing with uncertainty is totally up to them. This is the main reason for the users getting lost in the hypertext environment. Here, the uncertainty appears as follows:
   (a) The users have to find out their locations in the information web;
   (b) The users have to decide by themselves which links to follow.

4. Very little research has been done on this area. That seen in the literature includes Frisse's use of Bayesian network in highly-structured hypertext [67] and Pausch & Detmer's experiment of using node popularity as a hypertext browsing aid [121]. Whereas the former is not applicable to the general hypertext environment, the latter has the danger of misleading as the logical, or semantic relationship behind the link structure should always be the most important motivation for browsing.
7.2.3 Soft-link Hypertext

As the soft-link hypertext model uses a conceptual index (global + local) structure in its Index Space, it has the characteristics of traditional search and retrieval and that of traditional hard-link hypertext. Uncertainty in the soft-link hypertext model appears as follows:

1. As in the traditional hypertext model, search formulation should not cause much uncertainty in the soft-link hypertext environment;

2. The automatic formulation mechanism for the conceptual index (Chapter 5) could minimize the uncertainty involved in the hyperization process [Section 7.2.2] by facilitating systematically a data structure, presenting all options to the user, and making these hypertext options explicit and transparent;

3. In the soft-link hypertext model, searching (browsing) process is the main factor in which uncertainty is caused. Here, uncertainty is presented as which soft link(s) the information user should follow in order to meet the information needs.

It will be shown next that some mechanisms are accommodated in the soft-link hypertext model to offer on-line help and guidance to the user in dealing with uncertainty in the soft-link hypertext environment. These auxiliary tools are adopted to help the information user to deal with uncertainty during browsing in an Information Space.

1. In the soft-link hypertext model, the conceptual index is more explicit and transparent than the link structure used in the traditional hard-link hypertext, the uncertain problem can be alleviated by the user understanding the data structure better, and playing a more active role in information searching and browsing. Furthermore, as the conceptual index provides both global and local access to any concepts and enhances the user's mobility in an Information Space, it can help the user to avoid the confusion or getting lost caused by mere browsing locally in the hard-link hypertext environment. In other words, even if a user loses the orientation, it should be easier for him/her to trace the browsing trail and get re-oriented;

2. In the soft-link hypertext, the searching process is more transparent, and the information user has more interactions with the system during the information searching and browsing. Uncertainty can be alleviated by the information user having a better understanding of, and becoming more involved in, the searching
process. Meanwhile, the soft-link hypertext system provides improved facilities for its users via a superior user modelling mechanism;

3. In the soft-link hypertext model, various sophisticated semantic networks, including thesaurus and Bayesian networks [139] [67], could be adopted easily in the conceptual index to provide on-line help. Furthermore, powerful intelligent mechanisms are used in searching process to provide on-line navigation guidance during information search and retrieval. These mechanisms constitute help in the following three approaches:

- General information on which soft link(s) are more popular is provided as guidance for local browsing (LTM);
- Past browsing experiences (trails) are recorded and used, as recommendation for later users (LTM);
- An information user's previous browsing trail is used as basis for the reasoning mechanism to make predictions of the user's next move (STM). Thus, a short list of soft-link options could be tailored for each individual user.

From the next section, these intelligent mechanisms, including user modelling, learning and reasoning are discussed in detail.

7.3 User Modelling in Soft-link Hypertext

7.3.1 User Modelling

User modelling is an important feature for information systems. By building a good model of the user, i.e., establishing some picture of the information user's background, domain knowledge, familiarity with the system and the search goals, an information system can tailor its performance for the individual user.

However, although user modelling is very important for the success of information search and retrieval, reports about user modelling study are very rare in the IR literature. Generally speaking, user modelling mechanisms adopted in the present IR systems are very crude; they are often limited only to basic man-machine interactions.
1. Most existing IR systems gather user information by conducting a dialogue with the user through simple man-machine interactions. The advantages of such an approach are:

- This mechanism is well defined;
- It is easy to implement;
- Some limited changes and adjustments to the initially established search goals can be made via user feedbacks.

However, such a mechanism often results in a poor-quality user model. Its main disadvantages include:

- The collected information may be inaccurate, which often leads to uncertainty in IR. The information user might not know the answer to the questions asked by the system, or find it difficult to answer. The worst situation is that a user might become so frustrated by the questions that he/she could give up before a session starts;
- A well-developed questioning process is needed in such kind of systems. This, nevertheless, is an upside-down procedure. In other words, developing such a questioning device requires the designer to anticipate all the potential answers, which could be very difficult;
- The built-up user model is only a canonical, static and discontinuous model. Moreover, in the cases where the user has more than one inquiry, these multiple queries are often treated as unrelated with each other. This leads to the loss of associative information.

2. Hypertext should be an ideal environment for formulating a good user model because of its nature of supporting the recognition and browsing:

- Hypertext systems often employ a sophisticated and well designed user interface, and provide a pleasant environment for man-machine interaction;
- It is the defining element in hypertext to follow human associations, and to organize and retrieve information through these machine-supported associations. Hypertext should naturally be more user-centered than traditional IR models;
- The information user is usually able to express the information needs better because the man-machine interaction in hypertext environment relies more on the recognition than the recall;
• The information user has more interactions with the system during the browsing process, such as choosing hypertext anchors and links, searching forward or backward, etc.; moreover, the information-seeking is considered as a dynamic process;

• Generally speaking, the information-seeking process in the hypertext environment is a longer and more continuous process, during which a better user model could be obtained.

Nevertheless, the study of user modelling in hypertext has proven to be rare. Until now, little has been reported about this aspect of hypertext.

Ideally, a user model should be inferred from the information user's responses to the system, i.e., his/her information-seeking behaviour during the information retrieval process. This makes more demands on intelligence from the system, which is exactly what the soft-link hypertext model could provide in its Index Space. As one of the sub-hypotheses examined in this research, a better user modelling mechanism is provided in the soft-link hypertext model. The mechanism is set out below.

7.3.2 User Modelling in Soft-link Hypertext

As discussed in the previous section, hypertext should be an ideal environment for a good user modelling mechanism. The soft-link hypertext model has enhanced this property further, as even more man-machine interactions are involved in operating run-time soft links. In this section, a user modelling mechanism used in the soft-link hypertext model is presented. Its main characteristics include the following:

1. It supports both canonical and individual user modelling;

One side of the user modelling mechanism supported by long-term memory (LTM) assigns the users to a certain predefined subset of user groups according to its understanding of the users and provides services based on the general knowledge about the canonical groups; another side of it is supported by the short-term memory (STM) and records each individual user's information-seeking trail so as to treat him/her as a separate individual.
2. It supports both explicit and implicit user modelling;

The explicit model lets the information users provide their own models, while in the implicit model, the user model is inferred or abstracted by the system on the basis of the information users' behaviours. Considering the potential difficulty of implicit modelling, Daniels proposed "The best approach... is to allow the system to form a reasonable initial user model, perhaps based on the user's answer to a few preliminary questions, or on known characteristics of the overall user community - i.e. on a limited amount of whatever information is available to, or elicited by, the system. As a user proceeds to interact with the system, further information will become available, and the system can begin to update its user model, revise its hypothesis, alter default values, etc." [48].

This procedure is exactly how the user modelling mechanism operates in the soft-link hypertext model. Here, explicit user modelling is supported in conventional man-machine interfaces (windows, menu, etc.); more importantly, implicit user modelling mechanism is adopted by updating activation level of the concept in the Information Space after each interaction during the whole information-seeking procedure.

3. The user modelling is a long-term and continuous process;

In the soft-link hypertext model, all queries are treated as related to each other so that the information-seeking process during an IR session is considered as a long-term and continuous process. The earlier queries are collected, as $PATH^*$ in STM, and are used as basis for the system to reason for later movements.

4. It supports dynamic user modelling;

Dynamic user modelling is supported in the soft-link hypertext model when the information user's search goals alter during the session, either locally and globally. The activation spreading process can catch and reflect this change.

In the soft-link hypertext model, the user model is accommodated by a n-dimension vector. Each dimension of the vector represents an individual concept in the Information Space, and it is measured by a numerical pointer ($a_i^*$) called activation level. As defined in Chapter 6, for any $V_{Ci} \in V_C: 0 \leq a_i^* \leq 1$.

$$a^* = [a_0^* \ a_1^* \ ... \ a_n^*]$$
Before a user starts an IR session, all elements in the vector are set to zero. Therefore, 
$[0, 0, ..., 0]$ represents the fact that the system's knowledge about its user is zero.

While the session goes on, the user constantly interact with the system in his/her information seeking and browsing process. As a result, this vector, $[a_0^*, a_1^*, ..., a_n^*]$, is constantly updated by following the activation rules defined in Section 6.4. Such an activation spreading methodology is adopted by the system to update its knowledge about the user's information interests. This process represents the user modelling procedure employed in the soft-link hypertext model.

For example, the activation vector may present the following pattern:

$$[1 \ 0.2 \ ... \ 0.8 \ 0 \ ... \ 0.4]$$

This vector represents a system's specific user model at one given moment. Each number represents the probability that the related concept is relevant to the user's information interests. The larger $a_i^*$ is, the more relevant the concept $V_{Ci}$ is to the user's information needs.

1. For a concept the user is interested in — $a_i^* = 1$;
2. For a concept the user may be interested in — $0 < a_i^* < 1$;
3. For a concept the user is not interested in — $a_i^* = 0$.

### 7.4 Learning in Soft-link Hypertext

Learning is a very important feature for the soft-link hypertext model. It reflects the evolvement and change that occur in the behaviour of the model as a result of experiences over time. In the soft-link hypertext model, learning represents the process of tuning the weights of soft links to reflect the general popularity of the links to which they are related; it also records successful browsing trails ($PATH^*$).

Learning provides the knowledge base for user modelling and reasoning mechanisms in the model.

Learning in the soft-link hypertext model includes *long-term memory* (LTM) and *short-term memory* (STM). The difference between them, defined in neural networks [6], is that STM contains items that may be lost if one is distracted, whilst LTM contains items retained from some earlier period and lasts much longer.
7.4.1 Long-term Memory

In the soft-link hypertext model, LTM stands for the changes that occur in the behaviour of the model over a long time span, as a result of its experience over different Document Spaces and information users. In other words, LTM reflects the changes in two aspects: an Information Space and its applications. LTM records the following information:

1. All information components stored in the Document Space;
2. The conceptual index in the Index Space;
3. The weights associated with connections in the conceptual index;
4. Successful browsing trails by different users, \{$PATH^u \mid u = 0, 1, \ldots$\}.

The LTM learning mechanism, which is mainly based on the backward learning process in the connectionist network in the Index Space (discussed in Chapter 6), includes:

1. Adding/deleting concept(s) in the Information Space [Section 6.3];
2. Adding/deleting connection(s) in the Information Space [Section 6.3];
3. Increasing/decreasing the weights of connections (self-adjustment) [Section 6.5];
4. Increasing/decreasing the weights of connections (because of changes of environment, e.g., the special authorized soft links etc.) [To be discussed];
5. Recording different successful information-seeking trails, $PATH^u$, at the end of each IR session (Chapters 4 & 6) [Section 7.5].

7.4.2 Short-term Memory

STM represents the adaptation of a system's behaviour for a specific user during one IR session. Although STM is only valid for one session, it records the present users' user models, follows their information-seeking trails ($PATH^*$) and provides the direct information services to the users; therefore, it presents the most important feature of the model. STM records the following information:

1. The present user's user model, $a^* = \{a_0^*, a_1^*, \ldots, a_n^*\}$;
2. The present user's information-seeking trail, in the form of $PATH^*$. 
The learning process for STM includes:

1. Recording the user’s information needs, which is the result of the activation vector, \( a^* \), at the end of each spreading phase [Section 6.4];

2. Recording the user’s information-seeking trail (History), in the form of \( PATH^* \) [Section 6.4].

### 7.4.3 Relation between STM and LTM

The relation between STM and LTM are:

1. STM is the system’s memory during one IR session, whereas LTM is the memory which should last for the whole life-span of a system;

2. STM remembers
   - The present user model, \( a^* = \{ a^*_0, a^*_1, ..., a^*_n \} \);
   - The present user’s information-seeking trail (History), in the form of \( PATH^* \),

   whereas LTM records:
   - All documents and the conceptual index in an Information Space;
   - The weights associated with those connections;
   - Different users’ successful browsing trails \( \{ PATH^u | u = 0, 1, ... \} \).

3. At the end of an IR session, the browsing trail stored in STM, \( PATH^* \), becomes the newest information to be stored in LTM in the form of \( PATH^u, (u = 0, 1, ... ) \).

4. LTM might fade with time, which is represented in the model as the following:
   - As the competitive learning is used for LTM (Section 6.3), a soft link becomes prominent by the difference between its weight and those of others in the same layer. If the weights of the connections in a Kohonen layer increase the same amount, the relative positions of these links remain the same. This evenly distributed increase is considered to fade away;
   - Because of the limitation of machine memory, only a limited number of browsing trails can be recorded in LTM in a system. Consequently, when a new trail is recorded, an old one may be forgotten (fade away) if it has not been enhanced.
7.5 Reasoning in Soft-Link Hypertext

Definition 7.1 Formally, a browsing trail $PATH^*$ can be noted as a $1 \times n$ matrix:

$$P^* = [p_0^* \ p_1^* \ \ldots \ p_n^*]$$

where,

$$p_i^* = \begin{cases} 1 & \text{if the concept } V_{C_i} \in PATH^* \\ 0 & \text{otherwise} \end{cases}$$

Obviously, for any IR session, "$a_i^* = 1$" is used as a threshold for any concept $V_{C_i}$ to be included in $PATH^*$. It is, in fact, in accord with the fact that, if and only if a concept $V_{C_i}$ has been successfully visited by the user, the probability that it is relevant to the his/her information needs is 1. Here, the system ascertains that $V_{C_i}$ is relevant to the user's information interests.

Definition 7.2 Similarity between two trails, $PATH^u$ and $PATH^v$, is defined as:

$$sim[P^u, P^v] = P^u \times [P^v]^T$$

$$= [p_0^u \ p_1^u \ \ldots \ p_n^u] \times \begin{bmatrix} p_0^v \\ p_1^v \\ \vdots \\ p_n^v \end{bmatrix}$$

$$= \sum_{i=0}^{n} p_i^u \times p_i^v$$

The function of the reasoning mechanism in the soft-link hypertext model is to update the probability that a concept $V_{C_i} \in V_C$ is judged as relevant to the user's information needs. The update occurs every time the user's interacts with the system. This reasoning procedure is based on the following information:

1. The user's past browsing trail, $PATH^*$;
2. The user's latest interaction, $V_{C_i}$;
3. Finally, the model's past experiences, stored in LTM.
\[ P\{V_{cj} \cap [V_{ci} \cup PATH^*]\} = P(V_{cj} \cap V_{ci}) + P(V_{cj} \cap PATH^*) \]
\[ = P(V_{cj}|V_{ci}) \times P(V_{ci}) + P(V_{cj}|PATH^*) \times P(PATH^*) \]
\[ = P(V_{cj}|V_{ci}) + P(V_{cj}|PATH^*) \] (2)

- where,
  
  \( s \) stands for the present IR session;
  
  \( V_{ci} \) is the concept the user chooses as the information request;
  
  \( PATH^* \) is the browsing trail the user has visited;

\( P\{V_{cj} \cap [V_{ci} \cup PATH^*]\} \) is the prediction of the probability that a concept, \( V_{cj} \in V_C \), is relevant to the user’s information needs. In fact, it is \( a_j \).

Therefore, from (2),

\[ a_j = P(V_{cj}|V_{ci}) + P(V_{cj}|PATH^*) \] (3)

1. The first part of right hand side of Equation (3), \( P(V_{cj}|V_{ci}) \), means the conditional probability that \( V_{cj} \) is relevant to the user’s information needs given that \( V_{ci} \) is. As \( V_{ci} \) is the concept the user clicks, the first part represents the result of activation spreading, which has been discussed in Chapter 6;

This is the reasoning process in which the present information query is considered as the basis for the next move.

2. The second part of right hand side of Equation (3), \( P(V_{cj}|PATH^*) \), stands for the conditional probability that \( V_{cj} \) is judged as relevant to the user’s information needs given that the past browsing trail is \( PATH^* \). It is defined as:

\[ P(V_{cj}|PATH^*) = \begin{cases} 
\lambda \times (1 - e^{-\text{Max}_{(\text{sim}(P^*, P^i))}}) & \text{if } V_{cj} \in P^i \ (i = 0, 1, \ldots); \\
0 & \text{otherwise}
\end{cases} \]

It represents the reasoning process in which the user’s past browsing trail and associations are considered as the evidence for making predictions.
The procedure is:

1. $PATH^* = PATH^* + \{V_{ci} \} + \{V_{cj}\}$, which is recorded in STM;

2. All browsing trails $\{ PATH^u | u = 0, 1, \ldots \}$ in LTM are compared with $PATH^*$ so that those which are most similar to $PATH^*$ are chosen as references;

3. The prediction of the probability is based on those references, as well as the similarity between $PATH^*$ and them;

4. $\lambda = \frac{2}{d^2(i)}$.

With (3), the soft-link hypertext model keeps on making predictions on the probability that a concept in Information Space is judged as relevant to the user's information needs. The reasoning process is based on the model's long-term memory (LTM) and short-term memory (STM). Two sets of auxiliary information are used:

1. Which soft link the user might use, judged on popularity of each link;

2. Which soft link the user might use, judged on basis that some previous users who shared the similar the information interests (browsing trails) with the user.

These predictions are used as the interacting mechanism, in the form of a short list of soft-link options, or charters, between the information user and the IR system in order to speed up the information-seeking process. Here, the schema ForwardSpreading (Section 6.4) should be rewritten as the following:

```
ForwardSpreading
\Delta IndexSpace
PutInOrder: (Concept \times \text{decoder}[A_C]) \rightarrow \text{seq Concept}
PATH^*: \text{IF CONCEPT}
V_{ci}? : \text{CONCEPT}
rep!: \text{seq Concept}
```

```
V_{ci}? \in V_C: a_i' = a_i + 1
\forall V_{cn} (V_{cn} \in V_C \land V_{cn} \notin PATH^*):
    a_n' = a_n + \sum w_{mn} \times a_m + \lambda \times (1 - e^{-\text{Max}([sim(P^*, P')])})
\forall V_{cn} \in PATH^*: a_n' = 1
PATH^* = PATH^* + \{V_{ci}?\}
rep! = \text{PutInOrder} (\text{Concept } \times \text{decoder}[A_C])
```
7.6 Intelligence in the Soft-link Hypertext Model

The intelligent mechanisms in the Index Space, which includes automatic index formulation, user modelling, learning and reasoning mechanisms, provide the main intelligence in the soft-link hypertext model. Their operations help to build up better communication between the information system and information user, and provide guidance for an efficient information retrieval in an uncertain environment.

With these intelligent mechanisms defined, it is now possible to put all elements together to formulate the soft-link hypertext model. Figure 7.1 shows such a model in an IR environment: the user accesses an Information Space in order to obtain some required information; the Information Space, represented by the soft-link hypertext model, is designed to make this search and retrieval process efficient. It can be seen that the model consists of three main elements:

1. The Document Space;
2. A conceptual index;
3. Four auxiliary intelligent mechanisms.
Whilst the useful information is in the Document Space, the conceptual index defines the state-space of the model, and these intelligent mechanisms control the dynamics and applications of the model. Nine schemas are adopted for these mechanisms, which are summarized in Appendix A.

<table>
<thead>
<tr>
<th></th>
<th>Index formulation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>InitIndexSpace</td>
<td>1. InitIndexSpace</td>
</tr>
<tr>
<td></td>
<td>AddConcept</td>
<td>2. AddConcept</td>
</tr>
<tr>
<td></td>
<td>RemoveConcept</td>
<td>3. RemoveConcept</td>
</tr>
<tr>
<td></td>
<td>AddConceptualRelation</td>
<td>4. AddConceptualRelation</td>
</tr>
<tr>
<td></td>
<td>RemoveConceptualRelation</td>
<td>5. RemoveConceptualRelation</td>
</tr>
<tr>
<td>2</td>
<td>User modelling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SessionStarts</td>
<td>1. SessionStarts</td>
</tr>
<tr>
<td>3</td>
<td>Reasoning</td>
<td>2. ForwardSpreading</td>
</tr>
<tr>
<td>4</td>
<td>Learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BackwardLearning-1</td>
<td>1. BackwardLearning-1</td>
</tr>
<tr>
<td></td>
<td>BackwardLearning-2</td>
<td>2. BackwardLearning-2</td>
</tr>
</tbody>
</table>

7.7 Summary

In this chapter, we have shown how intelligent mechanisms in the soft-link hypertext model can be developed and used to help the user retrieve information efficiently in an uncertain IR environment. From the system's viewpoint, the whole procedure is presented as making prediction, adjusting the user model and learning, making prediction again, adjusting and learning again ... By making predictions based on LTM and STM, a short list of anchors can be chosen from all the soft-link options; this short list may include the user's real information need and speed up this search and retrieval process. In addition, adjusting the user model continuously could help the system obtain a better understanding of its user's information interests; learning could enable the system accumulate the knowledge about the general user groups, as well as about each individual user. With these intelligent mechanisms, different elements have been finally combined to facilitate the soft-link hypertext model. The model, as well as its implementation, are assessed and evaluated fully in Chapter 8 and 9.
Chapter 8
An Intelligent Information System

The structure and different mechanisms of the soft-link hypertext model were developed in the previous chapters; the full model is presented and evaluated in this chapter. The overview of its structure, the features it provides and the improvements it enables are demonstrated. The chapter concludes with the description of an implementation of the soft-link hypertext model — the Enhanced SuperBook.

8.1 Overview of the Model

The soft-link hypertext model, as shown in Figure 8.1, is the main topic of this dissertation. Its structure includes three main elements: a Document Space, a conceptual index and some auxiliary mechanisms in its Index Space (Section 7.6).

![Diagram of the soft-link hypertext model vs. the Dexter model.]

Figure 8.1: The soft-link hypertext model vs. the Dexter model.
It can be seen from Figure 8.1 that such a soft-link hypertext model constitutes a standard Dexter model (Section 3.4.1). The main elements of the model are summarized as follows:

1. The Document Space — Within-Component Layer;

   The Document Space includes the whole collection of information components (documents) in an Information Space, which is the main source of useful information for the information user. All information components in the soft-link hypertext model are presented in their original full-text forms. In other words, the contents and structure within components (Within-Component Layer) inherit directly from those of original documents.

2. The Index Space — Storage Layer;

   The Index Space includes a conceptual index and all auxiliary mechanisms to support the operations and applications of the model (Figure 8.2). Its function is to provide supplementary information and tools, so as to make on-line information search and retrieval more efficient. The Index Space is the core of the soft-link hypertext model and the main contribution of this research.

![Figure 8.2: The structure of the Index Space](image-url)
It is the Index Space that distinguishes the soft-link hypertext model from other IR models. Whilst its main structure and operations have been represented in detail in Chapters 5, 6 and 7, the most important characteristics of the Index Space can be summarized as:

- Automation;
- User centered;
- Intelligent behaviour;
- Learning and evolution.

3. Presentation of the soft-link hypertext structure — Runtime Layer.

Although it is not defined as a part of the model, the presentation of soft links during the user-system interaction in information retrieval is very important for any implementation. This is discussed in more detail later in the chapter.

8.2 The Soft-link Hypertext Model — Description and Novel Attributes

In this section, the soft-link hypertext model is compared and contrasted with the standard Dexter hypertext reference model. The goal of this analysis is to ascertain the following points:

1. What the soft-link hypertext model is;
2. The difference between the model and other hypertext models;
3. The possibility of interchange and interoperability between the soft-link hypertext model and others.

8.2.1 The Dexter Hypertext Reference Model

The Dexter model is an attempt to capture, both formally and informally, the characteristics found in a wide range of existing and future hypertext systems. It provides a standard hypertext terminology coupled with a formal model of the important abstractions commonly found in a wide range of hypertext systems. Thus, the Dexter model serves as a standard against which to compare and contrast the characteristics and functionality of various hypertext (and non-hypertext) systems. The Dexter
model also serves as a basis on which to develop standards for interchange and inter-operability among hypertext systems [81] [82].

Figure 8.3: The three-layer Dexter model.

Figure 8.3 shows a depiction of the Dexter model as embedded in an actual hypertext system. As described in Section 3.4.1, the model divides a hypertext system into three layers: the Run-Time Layer, the Storage Layer and the Within-Component Layer. The main focus of the model is on the Storage Layer, which models the basic node-link data structure that is the essence of hypertext.

The Storage Layer describes the structure of a hypertext as a set of components:

1. Atomic components: which are what is typically thought of as "nodes" in a hypertext system. Although the real content of information contained in the atomic components exist outside the Storage Layer in the Within-Component Layer (Figure 8.3), the atomic components constitute the primitive in the model;

2. Links: entities that represent relations between different components;

3. Composite components: which are constructed out of other components;

4. Anchors: mechanisms to support the span-to-span hypertext structure.

In the Dexter model, each component is assigned a global unique identifier (UID). An anchor has two parts: an anchor identification (id) and an anchor value, so that it
can be uniquely identified across the whole Information Space by a component UID and its anchor id pair. A link is a sequence of 2 or more specifiers, which identifies the end anchors of the link, as well as its directions.

In addition to a data model, the Storage Layer defines a small set of operations that can be used to access and modify a hypertext system. All of these operations are defined in a way that they maintain the invariants of the model. These operations include:

1. Adding a component (atomic, composite), or a link;
2. Deleting a component, or a link;
3. Modifying the contents of a component, an anchor, or a link;
4. Mapping from a component's UID into the component's itself (the accessor function, i.e., the traditional retrieval process);
5. Mapping from a component specification to its UID (the resolver function, i.e., the traditional search process).

The Dexter model provides a standard hypertext terminology coupled with a formal model of the important abstractions that are commonly found in the present hypertext systems. Therefore, by comparing the soft-link hypertext model with the standard Dexter model, the answers to the following questions should be found:

1. Does the soft-link hypertext model constitute a hypertext model?
2. What similarities does it share with a wide range of existing hypertext systems?
3. What differences does it facilitate?
4. How does it improve information retrieval in hypertext?

8.2.2 The Soft-link Hypertext Model

The analysis of the soft-link hypertext model starts with drawing a figure similar to Figure 8.3. Figure 8.4 illustrates the model as embedded in an actual soft-link hypertext system. By comparing these two figures (Figures 8.3 and 8.4), the following points can be concluded:
Figure 8.4: The soft-link hypertext model, redisplay of Figure 8.1.

1. The Document Space in the soft-link hypertext model is identical to the Within-Component Layer in the Dexter model;

2. The state-space and the operations defined in the Index Space are very similar to those of the Storage Layer, with the following difference:
   - The atomic component in the former is reduced into the individual concept;
   - Each atomic component still has a UID associated with it, but here, the accessor function is not a one-to-one function anymore. It is a onto function. In other words, in the soft-link hypertext model, the accessor function may map an atomic component (a concept) to more than one document;
   - The composite component represents the composite concept, which is constructed out of other components;
   - As the atomic component is reduced to a concept, an anchor in the soft-link hypertext model is equal to an atomic component;
   - The link in the soft-link hypertext model is still a sequence of 2 specifiers, which identifies two end anchors (concepts) and its direction. It connects two anchors (concepts) and reflects the relationships between them.
   - Finally, whilst all the operations in the Dexter model are still supported in the soft-link hypertext model, some more (including learning, user modelling and reasoning) are defined (Chapter 7);
3. As in the Dexter model, the soft-link hypertext model does not specify the format for its front end.

The conclusion from the above discussion is that the **soft-link hypertext model does constitute a hypertext model**. It is identical, in its fundamental structure, to most of the existing hypertext systems; it also supports the general operations of hypertext applications and properly maintains the invariants of hypertext.

The **principal difference** lies on the fact that, in the soft-link hypertext model, the atomic component is reduced into an individual concept. This modification would not affect the real content of information stored in the Document Space (Within-Component Layer), nor would it affect the operations and applications of the model in the hypertext environment; nevertheless, it does result in the following improvements:

1. The effort in defining and producing the atomic component, which has been shown as time-consuming and costly in Chapter 3, becomes unnecessary;

2. The effort in defining and producing the anchor, which is similarly expensive and tedious, becomes unnecessary;

3. Identification and production of the atomic component and anchor (the concept in this case) become a global indexing process, in which automation has long been introduced;

4. Furthermore, as the hypertext link structure becomes a **semantic relationship** among different concepts, hyperization in the soft-link hypertext model becomes building up the related semantic relationship. The latter should be easier because of the following reasons:
   - Thesauri and various kinds of term-relationship collections have long been studied and used in libraries and publications;
   - Applications of machine-supported semantic networks in information retrieval have been investigated in some limited environments [68] [139] [44];
   - An automatic formulation mechanism has been introduced in the soft-link hypertext model (Chapter 5) to identify and build up a special semantic network for hypertext.

5. Finally, control and manipulation of such a conceptual index for the soft-link hypertext model should be easier and more disciplined than that of the irregular link structure seen in most existing systems.
8.3 Features of the Model

From above discussions, it can be seen that the soft-link hypertext model adequately avoids some problems existing in the present IR systems, and provides great potential for efficient and effective information retrieval services. Whilst some of this potential has been explored in Chapters 5, 6 and 7, its main features can be summarized as follows:

1. A User-centered System.

   It is a user-centered model, because only guidance and suggestions and not controls are provided. All associative options are offered for the information user to make choices during the information-seeking process. Users may roam freely in the Information Space; more importantly, their thinking associations direct the run-time soft-linking process and control which link(s) to follow;

2. User Modelling.

   In the soft-link hypertext model, each information user is treated as an individual who might belong to one or more of canonical user groups. The information needs can either be explicitly expressed, or inferred implicitly from the user's information-seeking behaviours. Furthermore, the browsing process is considered as a continuous and dynamic process.

   This user modelling mechanism adopted in the model is useful for building up a more natural and complete user model, and enhancing the understanding of the system about its users;

3. Search and Retrieval.

   Recall and recognition are two equally important cognitive activities in human memory; they are needed at different phases during an information-seeking process. Mechanisms for facilitating both activities, i.e., the global index and local conceptual associations, are provided in the soft-link hypertext model. Here, the traditional IR and hypertext are integrated into one information retrieval infrastructure, which merges the benefits of both;


   One of the advantages of the conceptual index is that it could be formulated automatically. This avoids the tedious and costly hyperization process in the traditional hypertext environment. Whereas the global indexing is based on
traditional full-text indexing techniques, local associations are formulated by facilitating semantic relationships between each index term and other related terms. Should the natural language processing techniques be available, this local association formulation could have been performed better. However, at the moment, a set of non-syntactic & non-semantic formulation rules is employed to generate reasonably good associations. Furthermore, a learning mechanism is adopted in the model to improve these associations (soft links);


As the conceptual index provides the facility for local association, each of the connections is a potential hypertext link by invoking which the information user could be transferred from one component to another, or from one concept to another. This local associative structure, i.e., the soft links, represents the most important feature which distinguishes the model from the others. Although these run-time navigation and linking process may increase the user’s overload in making choices among the potential soft links, auxiliary tools are provided, via intelligent mechanisms adopted in the model, to provide a short list of soft-link options tailored to each individual user. This should alleviate the overload problem and speed up the information-seeking process.

Meanwhile, all features of traditional hard-link hypertext are still supported in the model, e.g., as specially authorized connections by the authors (editors), or an information user’s personalized information.

- Manually authorized hard links are facilitated by adjusting weights on the related links once they are marked. The conceptual structure and the auxiliary information in the Index Space can make this process semi-automatic, and therefore, much easier;

- Personal information management facilities, like footnotes, etc., are also supported, in form of personal soft path \( P A T H^* \) (See the next page).

The relationship between the soft and hard link is that the hard link is a subset of the soft link. Its weight is usually heavier than that of an ordinary soft link in the same layer, so as to imply the emphasis. The most popular soft links could become semi-hard links, or even hard links, once their weights are enhanced to levels with which they could permanently dominate the competition.

The concept of path, or ordered traversal of some links in a hypertext, has been a part of the hypertext notion from its early formation. In the soft-link hypertext model, not only are the paths set up purposely by authors or editors available for the information user, but also some of the previous users' browsing trails are potential paths, which could be followed by later users. The former is called hard paths, the latter soft paths. The soft-path mechanism is supported by STM and LTM in the model, in the form:

$$a_j(t) = P[V_{C_j}|PATH^*(t-1)]$$

Again, hard paths are a subset of soft paths, whose relative importance implies the emphasis. Personal information management in the soft-link hypertext model is supported by each information user's own soft paths.

7. Trail.

At any moment during an IR session, the information user's past browsing trail and all the visited concepts and links can be viewed, both globally and locally. The information is in STM: $$a^*(t) = \{a_1^*(t), a_2^*(t), ..., a_n^*(t)\}$$ and PATH*.

- It could be used to remind the information user of the concepts, links, or documents which have been visited;
- It facilitates the Goback mechanism for any steps.


Figure 8.5: The relationships between intelligent mechanisms and the user.
As distinct from other IR models, the soft-link hypertext model facilitates powerful learning and reasoning mechanisms, which follow human intelligence closely. The operations of these mechanisms (Figure 8.5) enable the model to have intelligent behaviours, accumulate the past experiences, improve its services, and finally, deal with information retrieval in uncertain circumstances.

8.3.1 Summary

In summary, mechanisms are developed in the soft-link hypertext model to generate and manipulate the data structure (the conceptual index), and to use it to make the information search and retrieval more efficient:

1. An index formulation mechanism is used to build up automatically the data structure. Although, presently, the quality of this formulation is affected by the limitation of natural language processing technologies, a reasonably good conceptual index can still be generated, which is further improved by self-adjustment and evolution;

2. With such a model, the user could benefit from the following three aspects:
   - The author and the editor get effective helps in hyperization;
   - The users have more options to choose in order to pursue their personal interests and to search and retrieve the information required. They do not necessarily need to follow the author's (or editor's) settings;
   - More machine intelligence is used to provide on-line guidance, and to make information search and retrieval easier and faster:
     A user modelling mechanism is used to infer the information user's needs implicitly so that the user could rely more on the recognition;
     Several methodologies are accommodated, including hard links, hard paths, soft paths and short lists of soft links, to reduce the user's overload, to speed up information-seeking process and to meet the requirements of different levels of information search and retrieval.

3. Finally, the model constantly adjusts and improves its behaviour via its learning and evolving mechanism.

In the next section, the model is evaluated in the information-seeking environment, whilst a version of its implementation is presented in Section 8.5.
8.4 Evaluation of the Model

The soft-link hypertext model provides an integrated infrastructure for information search and retrieval, which merges the positive aspects of different conventional IR features including the Boolean search, traditional hard-link hypertext and soft-link hypertext. The model is evaluated in this section, whereas the assessment of an implemented system of the model is given in Chapter 9.

8.4.1 Generation of Data Structure

As in the conventional IR environment, the soft-link hypertext model provides an automatic index formulation mechanism which is able to generate the conceptual index with the minimum human interference. This aspect of the model is superior to that in the traditional hard-link environment, where the whole data structure relies on the human hyperization process, and therefore, its applications are limited.

8.4.2 Applications of the Model in IR

Before starting to compare and contrast the soft-link hypertext model with other IR models, the information search and retrieval task is first divided into three categories. The purpose is to obtain a better definition of various IR environments. Each of the following environments specifies a special group of information needs.

1. Information Introduction: This represents the process in which the domain is new and unknown to the user; he/she searches for information as a novice reader. At this level, the user usually relies on the IR system to provide help and guidance during the information-seeking process. Here, the information retrieval task can be characterized as follows:

   - The user has difficulty in identifying exactly what is required;
   - The user is not able to express the information needs efficiently;
   - The user’s expectation of the search results is general, extensive, sketchy and introductory;
   - The search results should be generally on bibliographic, or document level;
   - Recall is more important than precision;
   - Passive interactions are more acceptable than active ones.
2. **Information Expansion:** This represents the process where the user is reasonably familiar with the domain and wants to expand his/her knowledge about some specific details in the domain. Its attributes include the following:

- The user knows exactly what is needed;
- The user might have difficulty in expressing the information needs;
- The user's expectation of the search results is more specific;
- Search outcome should be at document, section, or even paragraph level;
- Recall and precision are equally important.

3. **Information Pinpointing:** This represents the process in which the requested information is a specific term, concept, or a piece of information in the domain. The characteristics of this group of tasks can be expressed as follows:

- The user knows exactly what is required;
- The user's expectation of the search results is very specific;
- Precision is more important than recall;
- Search outcomes should be at the section, paragraph, or even sentence level.

In the following sections, different IR models are investigated in each of these three environments. The comparison of the soft-link hypertext model with others can be presented more clearly in such a well-defined setting.

### 8.4.3 Information Introduction

For this group, all sorts of menu-driven interacting IR systems (Gopher [76], World-Wide Web [11], etc.), guided tours (hard paths) and semantically-expandable systems [67][139][44] provide good solutions as the user relies on recognition and the system's promptness; with the conventional Boolean search, it is difficult to formulate good keyword(s) needed for the search and retrieval.

The user's recognition is strongly encouraged in the soft-link hypertext model. Clicking the recognized keywords constitutes the main user-system interactive formats, for both hypertext browsing and global searching. Furthermore, earlier users' successful browsing trails are provided, as soft paths, for the later users to follow. This mechanism provided in the soft-link hypertext model supplements the hard path in helping the user in Information Introduction phase.
8.4.4 Information Expansion

The user's main interests in Information Expansion are on expanding the information obtained. The information requested is reasonably well defined, and the specification for the search result is usually rigorous. For this group, traditional Boolean search, applied in combination with multiple attributes, presents an ideal model. This is the main stream of what is usually called information retrieval. With the help of feedback and probability theory adopted in a lot of more recent systems, it is usually very effective. The traditional hypertext model is inferior for this group of applications, because of its lack of global mobility and accessibility. If the requested information is, by coincidence, one of the destination anchors, then the user could retrieve it immediately. However, if it is not chosen as a destination anchor in some way, then, it is virtually impossible for the user find the information by browsing.

The role of the soft-link hypertext model in this category of information search and retrieval is very similar to that of the Boolean search. Here, the mapping from each individual concept to the related documents provided by the conceptual index can play a very important role.

8.4.5 Information Pinpointing

Part of the solution for this category of information search and retrieval has been provided in the highly-structured hard-link hypertext environments like dictionaries, encyclopedia and manuals; nevertheless, the main solution is still an open question. Here, the Boolean search is usually not accurate enough to pinpoint the information the user requests. The hard-link hypertext could provide the accessibility on the condition that a hypertext association has been linked to the destination information component. This is not always guaranteed.

Therefore, it is easy to see that the soft-link hypertext model has its unique role here for Information Pinpointing. The conceptual index adopted in the model provides local associations from simple concepts to composite concepts; it also supports direct access to the destination anchor which could be much more accurate than the Boolean search.
8.4.6 Summary

The comparison of these three IR models in different levels of information search and retrieval tasks can be summarized in Table 8.1.

<table>
<thead>
<tr>
<th>Application Environment</th>
<th>Conventional IR</th>
<th>Hard-link Hypertext</th>
<th>Soft-link Hypertext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Introductory</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Information Expansion</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Information Pinpointing</td>
<td></td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>

Table 8.1: Evaluations of IR models

8.5 Implementation of the Soft-link Hypertext Model

The soft-link hypertext model is implemented on SuperBook, which is a full-text information system. The implementation is in Tcl & Tk. Tcl & Tk is a simple yet powerful programming language for windowing applications [119]. The purpose of the implementation is to explore the possibility of practical application of the model, and have a demonstration of its features in a real IR environment. The Document Space used is the whole year's collection of chemical journals, Environmental Science and Technology (1991), from the American Chemical Society (ACS). The implemented new system is called the Enhanced SuperBook.

8.5.1 SuperBook

SuperBook is a full-text information retrieval system from Bellcore [56] [57]. It provides an elegant full-text browser with the efficient Boolean search mechanism, as well as a database of electronic version of ACS journals. SuperBook consists of three main elements:
1. A database, in this case, 10.2 MB\(^1\) of 1991 ACS journal *Environmental Science and Technology*, which is shown as ACS-DB in Figure 8.8;

2. A conventional full-text index (6 MB) for the Boolean search and all necessary tools for the browser, shown as IN-DB in Figure 8.8;

3. A powerful and user-friendly browser for information search and retrieval.

---

\(^1\)In this dissertation, MB means millions of bytes.
• The number of hits is shown during the searching process;
• The number of hits in the related component is shown in Table of Content;
• All hits are highlighted on Page.

3. Page. This is the window for representing the content of information components (Figure 8.7). Besides the text, footnotes, tables and chemical formulae are displayed on the right side of Page.

4. Auxiliary Buttons. Some auxiliary buttons are used in SuperBook:
• Following Page — for reading the next page;
• Previous Page — for reading the previous page;
• Search Forward — for searching the next hit (Figure 8.7);
• Search Backward — for searching the previous hit;

Figure 8.7: Searching for a word (e.g. polymer)
The overview of functional blocks in SuperBook is shown in Figure 8.8. The names of specific subroutines of SuperBook are presented in brackets. It should be noted that, in SuperBook, the feature of hypertext is so weak that it is identical to ordinary Boolean search function. Two I/O interface blocks are displayed in Figure 8.8, simply to specify the different processes for the same operation: when a user makes IR requests via *lookup*, the search interface is used; however, if a request is made in *page*, the SuperBook *hypertext* interface is used. This weak hypertext feature of SuperBook is modified in the Enhanced SuperBook; when a hypertext request is made, it is redirected to a much more powerful soft-link hypertext browser installed. This is further explained in Section 8.5.2.

![SuperBook I/O Interface](image)

Figure 8.8: The functional overview of blocks in SuperBook.

As presented earlier in the section, SuperBook includes a database of ACS journals (ACS-DB), a conventional full-text index (IN-DB) and a powerful browser (Figure 8.8). It offers an ideal testbed on which the soft-link hypertext model could be implemented, demonstrated and evaluated.
8.5.2 Implementation of the Model

Based on SuperBook, a soft-link hypertext system, called the Enhanced SuperBook, has been implemented.

Source code was written above the original SuperBook software. It accommodates a 46 KByte file, called sb.tcl, and was written in Tcl & Tk [119]. Whilst most of the original SuperBook subroutines remain unchanged, its hypertext interface was modified; all hypertext requests are redirected to a new hypertext browser installed in the Enhanced SuperBook. The extended functional blocks are shown in heavy-line boxes in Figure 8.9. They accommodate not only the original Boolean search function in the SuperBook, but also much powerful hard-link and soft-link features developed in the model (Chapters 5, 6, 7 and 8). The consequential hypertext browser supports various IR models in one infrastructure. The main functional blocks in the Enhanced SuperBook include a group of user I/O interfaces and three main subroutines: makeLink, useLink and impLink.

Figure 8.9: The functional overview of Enhanced SuperBook.
Thus, the improvement made in the Enhanced SuperBook is that all hypertext requests are redirected via the SuperBook hypertext interface to the hypertext browser developed in the thesis. Here, our implementation is able to make full use of the SuperBook features and successful extend its hypertext functionality to facilitate the soft-link hypertext model discussed in the thesis.

The SuperBook database of the ACS journals, ACS-DB, remains unchanged. Its full-text index (IN-DB) has been further extended to a conceptual index for the Enhanced SuperBook, which is stored in conceptualIN-DB. The former could well be removed; currently, it is retained for the ordinary Boolean search until more features of the conceptual index have been installed. At that time, the subroutine, wSearch, should search directly the conceptualIN-DB in the same way as it does now on IN-DB.

The three subroutines newly developed are explained as follows:

1. **makeLink** — formulation of the semantic network;

   The non-syntactic & non-semantic formulation rules introduced in Chapter 5 was fully installed in **makeLink**. Its function is to change the original SuperBook index and extend it to the conceptual index needed in the Enhanced SuperBook. **makeLink** checks each index term introduced by the SuperBook and finds out all the meaningful neighbours of the term in the ACS database. These formulated soft links are stored in a file, with the related term as its file name;

2. **useLink** — the soft-link hypertext browser;

   A soft-link browser is installed in sb.tcl as the main extension from the SuperBook to Enhanced SuperBook. The browser is composed of three layers of windows, each of which is tailored to a specific information services. More detailed information about the interface of the Enhanced SuperBook is given in Section 8.5.5. Each time a hypertext request is made, **useLink** spreads the activation level and presents the relevant information destination in the decreasing order of the estimated appropriateness to the user’s information needs.

3. **impLink** — the learning mechanism adopted in the model.

   The connectionist network learning algorithm discussed in Chapters 6 and 7 was fully implemented in sb.tcl. A weight is stored side by side with each soft link. It stands for the general popularity of the related soft link. The weights are automatically updated after each hypertext visit.
Enhanced SuperBook

As the result of the implementation, the Enhanced SuperBook has been in full operation. It accommodates the soft-link hypertext model and facilitates an integrated browser. As shown in Figure 8.9, three main elements in the Enhanced SuperBook are:

1. A database of 1991 ACS journal, *Environmental Science and Technology*, which includes 10.2 MB data and is inherited directly from SuperBook (ACS-DB);
2. A more sophisticated conceptual index (conceptualIN-DB), which was developed, by applying the automatic index formulation mechanism, on the full-text index provided in SuperBook. This mechanism is an implementation of the non-syntactic & non-semantic formulation rules (Chapter 5);
3. An extra interface for the integrated information service (Hypertext Browser in Figure 8.9), besides the original interfaces provided in SuperBook.

8.5.3 Databases

The Enhanced SuperBook includes the whole year’s collection of ACS journal, *Environmental Science and Technology* (1991), from the American Chemical Society (ACS). The domain is Chemistry. The whole database include:

1. Document Space, which takes 10.2 MB storage space (ACS-DB);
2. Index Space, which takes 12.3 MB storage space, including a global index, a semantic network and the weights related with each soft link (conceptualIN-DB);
3. A number of browsing trails (conceptualIN-DB).

Accepted as a global index, the full-text index provided in SuperBook (IN-DB) has been further extended into a conceptual index (conceptualIN-DB) by the automatic index formulation mechanism discussed in Chapter 5. The generated conceptual index becomes the main data structure for the Enhanced SuperBook. Whilst the original full-text global index (IN-DB) used in SuperBook could have been removed, it remains currently to support the ordinary Boolean search.
8.5.4 The Browser

A browser is built up in the Enhanced SuperBook, above the facilities provided by SuperBook, for browsing the soft-link hypertext (global index + local association). Once the soft-link hypertext mechanism is invoked by clicking a hypertext source anchor on Page, the soft-link hypertext browser is opened (e.g., polymer in Figure 8.10). It supports several features, including hard links, soft links, hard paths and soft paths. General usage instructions are given on the front page of the browser.

![Enhanced SuperBook](image)

Figure 8.10: Invoking the soft-link hypertext browser.

8.5.5 Interfaces

When the hypertext mechanism is invoked by clicking concept(s) in Page (polymer in Figure 8.10), the user is able to search and retrieve information via the soft-link hypertext browser. The interfaces for the browser is organized in a two-level hierarchical structure (Figure 8.11).
Figure 8.11: The interfaces of the soft-link hypertext browser.

1. Menu;
A number of features in the Enhanced SuperBook, which provides information services for various information-seeking models, can be chosen by choosing the related menu options. These options include:

- **Defaults (Figure 8.12);**
  This allows the user to follow the *hard links*, if any, which are pre-set from the present source anchor to another information component. These *hard links* can be set up by:
  
  (a) Authors of the document(s);
  
  (b) The user, for personal information management;
  
  (c) Other users, if the information is accessible.

- **Trails (Figure 8.13);**
  This allows the user to browse the Information Space via paths. The paths can be set up by:
  
  (a) Authors of the document(s), as *hard paths*;
  
  (b) The user, for personal information management;
  
  (c) Other users, as *soft paths*. 
Figure 8.12: Default — applications of the hard links.

- **Search**;

Combination Boolean searches could be facilitated when this option is chosen. The searchable attributes include:

(a) Title;
(b) Author;
(c) Keyword;
(d) Publisher.

- **Tailor**;

This is for a personally-tailored interface and information system, in which personal information about information background and special stopword requests could be inputted.
Figure 8.13: Trails — application of soft and hard paths.

- **History;**
This option provides the information user the opportunities to view the past information-seeking trail, i.e., the concepts, links, or components visited during the session, and backtrack any information components wanted.

- **Help;**
Help information would be organized in hypertext environment for the information user.

- **Cancel.**
Cancel is used to exit either the soft-link hypertext browser or the Enhanced SuperBook.
2. Question Document (Figure 8.14);

Beside using more traditional information-seeking approaches provided as options in Menu, the other main alternative is to use the soft-link hypertext mechanism, i.e., the global index and local browsing process, provided in the system. This option can be chosen by a user clicking Question Document and getting into the second level, Glossary and Context (Figure 8.11).

![Figure 8.14: Invoking the soft-link hypertext mechanism.](image)

- **Glossary**;

When the soft-link hypertext mechanism is chosen, all local associations related to the chosen concept are listed, in the decreasing order of their activation levels, on a window called Glossary (polymer in Figure 8.14). Choosing any one of these associations on Glossary (e.g. Polymer formation in Figure 8.15) transfers the user to the specified hypertext destination anchor in another information component (Figure 8.15).
Figure 8.15: Use of soft links (e.g. Polymer formation)

- **Context.**

If the chosen associations are included in more than one information component in the Document Space, the context of these components are displayed in an adjoining window called *Context*. For example, in Figure 8.16, after the soft-link hypertext browser is invoked by the user clicking *polymer* in *Page*, *polymer diffusion* is chosen next as the requested soft link. Here, there exist three different information components in the Document Space, which include the requested concept (*polymer diffusion*). As a result, the context of each of these three destination components is displayed in *Context* (Figure 8.16), for further choice.
8.6 Conclusion

The soft-link hypertext model, which has been developed in Chapter 5, 6 and 7, has been summarized and presented in this chapter. From its features, it can be seen that the model should be able to provide more versatile services to suit different information needs; more importantly, it should facilitate a faster and easier IR tool for the information author and user. This model has been evaluated theoretically in the chapter, which proves the hypothesis that a more sophisticated data structure (the conceptual index) can provide a good basis for integrated IR services; an IR model based on such a data structure can improve the features of the more traditional models. The system implemented, the Enhanced SuperBook, is further evaluated in three experiments in Chapter 9. A fourth experiment was planned, but later abandoned, due to reasons given in that chapter.
Chapter 9
Evaluation of the System

Evaluation of the soft-link hypertext model and the system implemented, the Enhanced SuperBook, are presented in this chapter. Whilst the former has been discussed in Section 8.4, the experiments in which the Enhanced SuperBook is tested and assessed at the practical level are discussed in detail here. Four experiments have been designed, each of which tests a different aspect of the soft-link hypertext system. The emphasis is on the results of these experiments, the analysis of the results, and the evaluation of the system based on these experimental results, so that the hypothesis examined in this dissertation can be proven in the applicational information retrieval environment.

9.1 Evaluation

Generally, evaluation is to assess the designs and test the systems to ensure that they actually behave as expected [53]. In this thesis, four experiments have been designed. The evaluation procedure is organized as in Figure 9.1.

```
Evaluation

Evaluation of the Model
(Discussed in Section 8.4)  Evaluation of the System

Experiment 1  2  3  4
```

Figure 9.1: Evaluation of the system

Here, evaluation is to assess the soft-link hypertext model and its implementation, the Enhanced SuperBook, so that the hypotheses examined in the dissertation could be proven at both the theoretical and practical levels.
9.2 Evaluation of the Model

The evaluation of the soft-link hypertext model has three main goals:

1. To assess the extent of functionality of the model;
2. To evaluate the features provided by the model;
3. To compare the model with other conventional IR models.

In the evaluation, the functionality of, and features provided by, the soft-link hypertext model are compared and contrasted with other conventional IR models at three different levels of information retrieval environment; this should enable the hypotheses in the dissertation to be assessed at the theoretical level. The evaluation of the model is presented in Section 8.4. The conclusion is that, with a more sophisticated conceptual index, the soft-link hypertext model can provide an integrated IR infrastructure, improve the features accommodated by other traditional ones and make information retrieval more efficient and effective.

9.3 Evaluation of the System

The evaluation of the system implemented, the Enhanced SuperBook, is established in a series of laboratory-based experiments, in which outcomes of the Enhanced SuperBook (ES) in different IR settings are tested and compared with those of the original SuperBook (SU). Although the idea of the soft-link hypertext model has been shown theoretically in the last several chapters to be a good model for information retrieval, especially for hypertext applications, a further assessment of it in some real IR environments should be valuable. The goals of the experiments are to measure the cost, effectiveness and benefits of the Enhanced SuperBook in information formulation, presentation, search and retrieval. This ensures the following points:

1. The strength and weakness of the model can be investigated further;
2. The hypothesis examined in this research can be proven practically;
3. Some concrete results and conclusions can be drawn from the trial and error;
4. Pointers can be provided to further developments in information retrieval.
Figure 9.2: The Flow Chart of Information Retrieval.
9.3.1 The Variables Measured

The evaluation of an information search and retrieval system can be carried out in many different ways, depending on the type of system considered and the viewpoint taken. A large number of different variables may, therefore, be considered [133]. In this research work, the following variables are used to evaluate the system:

1. **Recall**: Recall (\( R \)) is defined as follows:
   \[
   R = \frac{\text{Number of items retrieved and relevant}}{\text{Total relevant in collection}} \quad (\%)
   \]

2. **Precision**: Precision (\( P \)) is defined as follows:
   \[
   P = \frac{\text{Number of items retrieved and relevant}}{\text{Total retrieved in collection}} \quad (\%)
   \]
   (Depending on meaning of *relevance* and *collection*, recall and precision can stand at different levels and be measured on different viewpoints. This will be stated more clearly once these two variables are used in the experiments.)

3. **Response time**: Response time (\( RT \)) is defined as the time needed for a user-system interaction, i.e., from the moment an IR job is submitted to the system, to the moment the result is presented back to the user. These jobs include processes from Submission of Queries to Receive Search Results, and from Retrieve Documents to obtain the documents (not shown) in Figure 9.2. 
   
   \( RT \) is a measurement of a system's promptness in responding the user's request;

4. **Search time**: Search time (\( T \)) is defined as the total time the user needs to obtain a concrete piece of information with an IR system (from Formulation of Queries to Meet Information Needs in Figure 9.2). It can be seen that \( T \) usually consists of three fundamental elements, or the combinations of them:
   - Formulation of queries;
   - Several presentation time \( T \) for information search and retrieval;
   - Obtaining the information, or giving up.

   \( T \) is used to measure a system's effectiveness and efficiency in IR;

5. **User actions**: User actions (\( U \)) are defined as the number of IR jobs the user needs to fulfill in order to obtain a concrete piece of information.

   \( U \) is also used to measure a system's effectiveness and efficiency in IR.
9.3.2 Methodology

The overall methodology used in these experiments is the fundamental one popularly used in the scientific area [79]: every experiment includes several experimental groups, each of which represents a unique treatment ($X_i$). Random assignments ($R_i$) are allocated to each experimental group, from which a number of outcomes ($O_{ij}$) are collected. This procedure is further explained as follows:

Experimental group 1: $R_1 \ X_1 \ O_{11}, O_{12} \ ...$

Experimental group 2: $R_2 \ X_2 \ O_{21}, O_{22} \ ...$

... ... ...

where: $R_i$ indicates a random assignment;
$X_i$ indicates a unique experimental treatment, e.g., SuperBook (SU);
$O_{ij}$ indicates an observation or measurement of the outcome.

1. It could be either the same assignment ($R_i = R_j$), or several different but random ones ($R_i \neq R_j$), are fed into different treatment groups;
2. Each treatment represents an independent variable which is under test in the experiment;
3. Measurement of the outcomes (dependent variables) from the different experimental groups (treatments) are collected and compared to see whether different treatments produce different outcomes and why.

9.3.3 Design of the Experiments

Four laboratory-based experiments have been designed for the assessment of the Enhanced SuperBook, as shown on the next page. IR is a quite complicated procedure, in which many factors, like the user's psychological feeling, motivation, background and familiarity with the system (both hardware and software), influence the outcomes. While these factors are difficult to measure and control, some small-scale laboratory-based experiments are still manageable. From the mean values of the outcomes, it is hoped that some tangible results can be reached. These results should not only reveal some detailed facts about the system, but also present the trends and provide pointers to further research in the area.
<table>
<thead>
<tr>
<th>Ex No.</th>
<th>$P_i$</th>
<th>$X_i$</th>
<th>$O_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp 1</td>
<td>40 keywords from 1991 ACS journal (10.2 MB)</td>
<td>Indexing mechanism (SU)</td>
<td>1. Recall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indexing mechanism (ES)</td>
<td>2. Precision</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Efficiency of formulation</td>
</tr>
<tr>
<td>Exp 2</td>
<td>Subject group 1</td>
<td>Browser and Interfaces (SU)</td>
<td>1. Recall and precision</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Response time</td>
</tr>
<tr>
<td></td>
<td>Subject group 2</td>
<td>Browser and Interfaces (ES)</td>
<td>3. User actions</td>
</tr>
<tr>
<td>Exp 3</td>
<td>Subject group 1</td>
<td>SuperBook Browser</td>
<td>1. Information obtained</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Accuracy of retrieved information</td>
</tr>
<tr>
<td></td>
<td>Subject group 2</td>
<td>ES with learnt soft links</td>
<td>4. Time/actions for Ques. 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5. Time/actions for Ques. 2</td>
</tr>
<tr>
<td></td>
<td>Subject group 3</td>
<td>ES with learning mechanism</td>
<td>6. Time/actions for Ques. 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7. Time/actions for Ques. 4</td>
</tr>
<tr>
<td></td>
<td>Subject group 4</td>
<td>ES without learning</td>
<td></td>
</tr>
<tr>
<td>Exp 4</td>
<td>Subject group 1</td>
<td>Soft-link hypertext (ES)</td>
<td>1. Efficiency of retrieval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Effectiveness of retrieval</td>
</tr>
<tr>
<td></td>
<td>Subject group 2</td>
<td>Hard-link hypertext (ES)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subject group 3</td>
<td>Traditional IR (SU)</td>
<td>3. Psychological feeling</td>
</tr>
</tbody>
</table>

Experiment 1 — to evaluate the automatic indexing and formulation mechanism used in the soft-link hypertext system;

Experiment 2 — to evaluate the acceptability of, and the user-overload caused by, the soft-link hypertext browser;

Experiment 3 — to compare the soft-link hypertext system with a Boolean search system, and evaluate the role of intelligent mechanisms in IR;

Experiment 4 — to evaluate different IR models in an integrated environment:
1. Their roles at different levels of information retrieval are investigated;
2. The user modelling mechanism in the system is evaluated;
3. The impacts of machine intelligence on the information search and retrieval, as well as on the information user, during the different information-seeking procedure are assessed.
Figure 9.3: Design of the Experiments
9.4 Experiment 1 — Index Formulation

9.4.1 Purpose and Design

The purpose of the experiment is to test the efficiency and effectiveness of the automatic index formulation mechanism used in the Enhanced SuperBook (ES). Here, the comparison is with the indexing mechanism used in the SuperBook (SU):

1. Two different treatments (independent variables) were used in Experiment 1:
   - The indexing mechanism in the SuperBook;
   - The indexing mechanism in the Enhanced SuperBook.

2. The assignments, a random sample of 40 words from 1991 ACS journal Environmental Science and Technology, were used for both treatments;

3. The experiment was executed in the following way: the assignments (40 words) underwent both treatments, and the outcomes (dependent variables) from both treatments were collected and compared. The dependent variables included:
   - Recall;
   - Precision.

Theoretically, the time needed for indexing, and the memory space needed for storage of the data structures should also be included as dependent variables to test the cost (efficiency) of the indexing mechanisms. However, they were not considered in the experiment for the following reasons:

- The cost of an IR system can be divided into two parts: the maintenance of the system and its applications. The former includes the cost of indexing and maintaining the index structure. Although it is of primary importance for operations of any IR systems, it is not the main concern of this research work. Here, only the latter was considered;
- The cost of indexing and maintenance is difficult to measure in an experimental situation, because it is dependent on real implementation of a system and the hardware platform used. Obviously, implementation of a complete IR system is beyond the scope of this research;
- For SuperBook and the Enhanced SuperBook, both indexing and formulating the semantic net are executed off-line, and therefore, they would not affect the on-line search and retrieval (browsing) anyway.
The hypothesis examined in the experiment is that the tested indexing and formulation mechanism (mainly the non-syntactic & non-semantic formulation rules used in the soft-link hypertext model) can facilitate a conceptual index suitable for hypertext environment efficiently, and with satisfactory recall and precision rates.

9.4.2 Experimental Process

1. 40 sample terms were chosen from the 1991 ACS journal, Environmental Science and Technology, to test the indexing mechanisms. They were divided into 2 groups;
   - Group 1 (20 words, shown in Table 9.1) was chosen via a participatory design process: about 20 chemists voluntarily noted down the first word coming into their minds (the words they were most interested in);
   - Group 2 (20 words, in Table 9.2) was chosen via a statistical process, i.e., the fifth word on every sixth page in a chemistry dictionary was chosen.

2. For each of these 40 keywords, the total number of hits (case-insensitive) included in the 1991 Volume of Environmental Science and Technology was formulated first, by using the Unix command grep;

3. For each of these keywords, the number of hits retrievable by SuperBook was also collected;

4. Recall for SuperBook was calculated as \( \frac{3}{2} \);

5. Precision for SuperBook is obviously 100%, as each of the retrieved information is the keywords itself;

6. Recall and precision for the global index of the Enhanced SuperBook is exactly the same as those of SuperBook (Chapter 5);

7. 2 \times 2 chemists (domain experts) took part in the experiment to assess, from the system designer’s point of view, the recall and precision of the conceptual index formulated by the mechanism;

8. They went through each hit and the related sentence to check if the formulated phrase was relevant to the original information. Those which were not were noted.

9. The recall of the conceptual index structure was calculated as \( \frac{(3)-(8)}{(3)} \);

10. The precision of the second-level index structure was calculated as \( \frac{(2)-(8)}{(3)} \).
9.4.3 Experimental Results.

1. 20 words (Group 1) were chosen via a participatory design process, in which about 20 chemists took part (Table 9.1);

2. 20 words (Group 2) were chosen via a statistical process (Table 9.2);

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>agent</td>
<td>13.6</td>
<td>100</td>
<td>13.6</td>
<td>100</td>
<td>13.3</td>
<td>97.7</td>
</tr>
<tr>
<td>2</td>
<td>biology</td>
<td>48.9</td>
<td>100</td>
<td>48.9</td>
<td>100</td>
<td>46.7</td>
<td>95.5</td>
</tr>
<tr>
<td>3</td>
<td>bonding</td>
<td>72.2</td>
<td>100</td>
<td>72.2</td>
<td>100</td>
<td>61.1</td>
<td>97.4</td>
</tr>
<tr>
<td>4</td>
<td>carbon</td>
<td>33.8</td>
<td>100</td>
<td>33.8</td>
<td>100</td>
<td>31.9</td>
<td>91.8</td>
</tr>
<tr>
<td>5</td>
<td>chemistry</td>
<td>77.8</td>
<td>100</td>
<td>77.8</td>
<td>100</td>
<td>73.7</td>
<td>94.7</td>
</tr>
<tr>
<td>6</td>
<td>condensation</td>
<td>83.3</td>
<td>100</td>
<td>83.3</td>
<td>100</td>
<td>81.3</td>
<td>91.4</td>
</tr>
<tr>
<td>7</td>
<td>diamond</td>
<td>50.0</td>
<td>100</td>
<td>50.0</td>
<td>100</td>
<td>50.0</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>dioxide</td>
<td>73.9</td>
<td>100</td>
<td>73.9</td>
<td>100</td>
<td>73.9</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>fiber</td>
<td>53.8</td>
<td>100</td>
<td>53.8</td>
<td>100</td>
<td>52.4</td>
<td>98.7</td>
</tr>
<tr>
<td>10</td>
<td>film</td>
<td>25.5</td>
<td>100</td>
<td>25.5</td>
<td>100</td>
<td>22.5</td>
<td>85.7</td>
</tr>
<tr>
<td>11</td>
<td>gas</td>
<td>42.5</td>
<td>100</td>
<td>42.5</td>
<td>100</td>
<td>41.1</td>
<td>93.5</td>
</tr>
<tr>
<td>12</td>
<td>immune</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>inflammation</td>
<td>27.3</td>
<td>100</td>
<td>27.3</td>
<td>100</td>
<td>27.3</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>inhibitors</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>81.8</td>
<td>81.8</td>
</tr>
<tr>
<td>15</td>
<td>molecule</td>
<td>30.6</td>
<td>100</td>
<td>30.6</td>
<td>100</td>
<td>26.3</td>
<td>86.0</td>
</tr>
<tr>
<td>16</td>
<td>phenomenon</td>
<td>94.3</td>
<td>100</td>
<td>94.3</td>
<td>100</td>
<td>34.3</td>
<td>37.4</td>
</tr>
<tr>
<td>17</td>
<td>polymer</td>
<td>29.4</td>
<td>100</td>
<td>29.4</td>
<td>100</td>
<td>28.0</td>
<td>95.3</td>
</tr>
<tr>
<td>18</td>
<td>sensor</td>
<td>47.5</td>
<td>100</td>
<td>47.5</td>
<td>100</td>
<td>29.5</td>
<td>62.1</td>
</tr>
<tr>
<td>19</td>
<td>temperature</td>
<td>52.7</td>
<td>100</td>
<td>52.7</td>
<td>100</td>
<td>23.4</td>
<td>44.3</td>
</tr>
<tr>
<td>20</td>
<td>thermal</td>
<td>63.3</td>
<td>100</td>
<td>63.3</td>
<td>100</td>
<td>58.4</td>
<td>92.1</td>
</tr>
</tbody>
</table>

Table 9.1: Index terms in Group 1

3. Recall and precision used for SuperBook in Tables 9.1 and 9.2 can be explained as follows:

Dependent on the definition of node (i.e. the smallest unit of information in an IR system), the meaning of relevancy in IR can be divided into three levels:

(a) For a system whose node is the document (the highest level), recall and precision is the measurement of the system's ability to spot a relevant document, in which the keyword(s) requested are included;

(b) For the system like Superbooks, (whose node is the paragraph), relevancy means whether the keyword(s) requested are included in a related paragraph. Obviously, meaning of relevant is more rigorous than that in (a);
<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>SuperBook</th>
<th>Enhanced SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>1</td>
<td>absorption</td>
<td>81.1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>amorphous</td>
<td>83.1</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>cohesive</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>density</td>
<td>70.8</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>distribution</td>
<td>60.6</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>elution</td>
<td>68.4</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>frequency</td>
<td>82.9</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>hydration</td>
<td>52.9</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>joint</td>
<td>88.2</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>methyl</td>
<td>8.0</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>nucleation</td>
<td>79.7</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>physical</td>
<td>75.9</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>radius</td>
<td>80.6</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>resin</td>
<td>52.7</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>rubber</td>
<td>44.0</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>scattering</td>
<td>91.7</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>temperature</td>
<td>52.7</td>
<td>100</td>
</tr>
<tr>
<td>18</td>
<td>tracer</td>
<td>53.2</td>
<td>100</td>
</tr>
<tr>
<td>19</td>
<td>uranium</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>20</td>
<td>vinyl</td>
<td>40.0</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 9.2: Index terms in Group 2

(c) The lowest level is the sentence level. Recall and precision at this level represent the measurement of the system's ability to spot the keyword(s) requested in the related sentences. As a result, the recall and precision at this level should be much lower than those of (a) and (b) for the same indexing methodology.

In view of the characteristics of hypertext in this research, Level (c) is used here as the measurement in Tables 9.1 and 9.2. It should be noted that, if judged on Levels (a) or (b), the recall and precision of SuperBook should be both 100%.

4. Recall and precision for the Enhanced SuperBook are explained as follows:

(a) For the global index in the Enhanced SuperBook, the recall and precision rates are exactly the same as those in SuperBook [3(c)];

(b) For the conceptual index, recall and precision are lower than 4(a).
9.4.4 Analysis of Results

1. These two groups are randomly sampled so that they represent the general distribution of the keywords in the information space. This can also be seen in Table 9.3.

<table>
<thead>
<tr>
<th>The compared items</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decreasing value of recall (%)</td>
<td>12.8</td>
<td>11.9</td>
</tr>
<tr>
<td>Decreasing value of precision (%)</td>
<td>12.8</td>
<td>12.9</td>
</tr>
</tbody>
</table>

Table 9.3: Decreasing percentage of recall and precision

Both recall and precision rates degenerate after the non-syntactic & non-semantic formulation rules are applied on the global index to facilitate the semantic net needed in the conceptual index. Mean degeneration rates for both groups are shown in Table 9.3. The cross-group differences are 7% and 0.7%. This shows that the random sampling of both assignment groups is acceptable;

2. The recall rates for SuperBook and the global index of the Enhanced SuperBook are very low, mainly for the following reasons:

- The level of measurement used in the experiment is the lowest (sentence level). The SuperBook browser is not full-text based, but paragraph-based, i.e., any number of hits of an index term in one paragraph is only counted as one hit [57]. Therefore although, from the information retrieval point of view, the recall rate of SuperBook could be 100%, it is actually quite low when the sentence level measurement is adopted in the experiment;
- Furthermore, lack of natural language processing techniques limits the recognition of those hits when they are expressed as pronouns;
- Synonyms, as well as isomorphic expressions like water and $H_2O$, are ignored at the moment in both SuperBook and Enhanced SuperBook, though the research results by Savoy [139], Frisse [67], Lucarella [104] and Turtle & Croft [159] are ready to be adopted in both.

3. As the conceptual index formulation is dependant on the global index, the recall rate for the conceptual index of the Enhanced SuperBook is even lower than that of global index (Tables 9.1 and 9.2). It is affected not only by the same reasons as those listed in 2, but also by the imprecise non-syntactic & non-semantic formulation rules adopted in the formulation mechanism;
4. Compared with the recall rates of SuperBook and the global index in the Enhanced SuperBook, that of the conceptual index of the Enhanced SuperBook degenerates on average by about 12% (Table 9.3). This means that the decreasing value caused by the non-syntactic & non-semantic formulation rules is about 12%. This should be considered as acceptable. Should the recall rate for the global index be higher (i.e., if a genuine full-text index with about 100% recall rate is used as the global index), that for the second-level should be above 88%.

5. Precision for the global index should be 100% for a full-text indexed system. Precision rates of the index mechanism in the SuperBook and the global index in the Enhanced SuperBook are both 100%.

6. As the conceptual index includes more information on each of its index terms (a phrase, instead of a keyword as in 5), the idea of relevance becomes more rigorous and accurate. Therefore, here, precision rate represents a lower level parameter than those in 5, and is unsurprisingly lower than 100%. The precision rate for the second-level index in the Enhanced SuperBook is shown on Tables 9.1 and 9.2.

7. Compared with the SuperBook, the precision rate for the second-level index in the Enhanced SuperBook decreases averagely by about 12% (Table 9.3).
   - It can be seen that overall rate is satisfactory (> 85%);
   - It is especially good on domain-specialized nouns like bonding, condensation, etc, but not so good on general terms and adjectives (e.g. phenomenon, temperature, etc.). This is not worrying as domain-specialized nouns and phrases are exactly what a hypertext system is targeting (Figure 5.1);
   - The precision rates for any items related with chemical formulae (film, molecule, ...) are lower. How to facilitate chemical formulae and the relevant expressions with non-syntactic & non-semantic formulation rules still remains a question;
   - Furthermore, with the intelligent mechanisms adopted in the system (Chapter 8), the low precision rate for the conceptual index could be supplemented by self-adjustment and evolution. This is discussed further in Experiment 3.
9.4.5. Conclusion

Experiment 1 tests the efficiency and effectiveness of the automatic index formulation mechanism used in the Enhanced SuperBook. It shows that the mechanism could formulate automatically the identifiers needed for both global index and local association (soft links) with satisfying recall and precision rates (> 85%). The degeneration rates for recall and precision caused by the non-syntactic & non-semantic formulation rules are both lower than 15%, compared with those in SuperBook. This should be considered as acceptable. Moreover, these parameters could be further improved by adopting a more thorough set of formulation rules, natural language processing techniques, or some domain or linguistic knowledge (in the form of semantic networks) in the system. Meanwhile, the intelligent mechanisms provided in the Enhanced SuperBook could facilitate an effective supplement for the imperfect conceptual index structure, which is evaluated further in Experiment 3.
9.5 Experiment 2 — Soft-link Hypertext Browser

9.5.1 Purpose and Design

This experiment is to test, from the user's point of view, the acceptability of the soft-link hypertext browser implemented in the Enhanced Superbook, and more importantly, to evaluate the user's cognitive overload brought by the added features in the system. It provides a comparison with the same features of SuperBook.

1. Two different treatments (independent variables) were used in the experiment:
   - The browser used in the SuperBook;
   - The browser used in the Enhanced SuperBook;

2. Four subjects (domain experts) were used as information users in the experiment. They were randomly divided into two groups: Group A and Group B. Each group underwent both treatments.

<table>
<thead>
<tr>
<th>Treatment groups</th>
<th>SuperBook Browser</th>
<th>Enhanced SuperBook Browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Group A</td>
<td>A-SU</td>
<td>A-ES</td>
</tr>
<tr>
<td>Subject Group B</td>
<td>B-SU</td>
<td>B-ES</td>
</tr>
</tbody>
</table>

Table 9.4: Distribution of the subject in two treatments

3. Four groups of data were collected and compared (Table 9.4): the vertical comparison assesses the validity of the collected data, whereas the horizontal comparison is to assess and compare the features in the two treatments. The dependent variables are six criteria popularly used in IR [134]:
   - Recall;
   - Precision;
   - Response time;
   - User actions;
   - Presentation;
   - Collection coverage.

The hypothesis examined in the experiment is that the cognitive overload caused by the additional interfaces (soft-link hypertext browser) would not affect the user's efficiency of information retrieval. The user should feel as at ease in such a hypertext environment as with the ordinary Boolean search systems.
9.5.2 Experimental Process

1. 40 keywords chosen in Experiment 1 were used again in the experiment: 20 in Group 1 (Table 9.1) and 20 in Group 2 (Table 9.2);

2. Overload testing: this task investigates the additional overload brought by the additional windows adopted in the Enhanced SuperBook, in the following steps:
   (a) One keyword (in the order of Group 1 first, Group 2 second) was taken from the list. The subject was shown the chosen keyword. The current page was set to the first page in which the chosen keyword appeared;
   (b) The subject was then asked to look for some information related to the chosen keyword. The list of information requested is provided in Appendix B;
   (c) The subject used either SuperBook (the Boolean search), or the Enhanced SuperBook (the soft-link hypertext), to search for the information. The distribution of treatments and subjects is as Table 9.5;
   (d) The information retrieval session ended when the subject either found the information requested, or was convinced that the information was not included in the Information Space and gave up the session;
   (e) The time spent on the session, as well as the user actions needed in completing the search and retrieval, was collected;
   (f) The experiment goes back to Step (a), until all 40 keywords were covered;

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Subject 1 in Group A</th>
<th>Subject 2 in Group A</th>
<th>Subject 1 in Group B</th>
<th>Subject 2 in Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keywords in Group 1 (20)</td>
<td>SU</td>
<td>ES</td>
<td>SU</td>
<td>ES</td>
</tr>
<tr>
<td>Keywords in Group 2 (20)</td>
<td>ES</td>
<td>SU</td>
<td>ES</td>
<td>SU</td>
</tr>
</tbody>
</table>

Table 9.5: Distribution of treatments and subjects in Experiment 2

3. Measurement of the browser: this assesses, from the user's point of view, the two browsers and compares their different features;

Each subject went through the whole list of 40 keywords. Based on the searching (or browsing) experiences and the retrieved information, he/she examined the parameters of both browsers (recall, precision, response time, and presentation).
9.5.3 Experimental Results

1. The results of overload testing are listed in Tables 9.6, 9.7, 9.8 and 9.9.
   - Table 9.6 and 9.7 list search times spent on Group 1 and 2;
   - Table 9.8 and 9.9 list the user actions needed for Group 1 and 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>No. of items by SuperBook</th>
<th>No. of items by E. SuperBook</th>
<th>SuperBook</th>
<th>E. SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A-SU</td>
<td>B-SU</td>
</tr>
<tr>
<td>1</td>
<td>agent</td>
<td>43</td>
<td>21</td>
<td>0.17</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>biology</td>
<td>22</td>
<td>13</td>
<td>0.92</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>bonding</td>
<td>39</td>
<td>22</td>
<td>2.62</td>
<td>0.90</td>
</tr>
<tr>
<td>4</td>
<td>carbon</td>
<td>707</td>
<td>224</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>chemistry</td>
<td>452</td>
<td>158</td>
<td>0.82</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>condensation</td>
<td>70</td>
<td>54</td>
<td>2.13</td>
<td>2.22</td>
</tr>
<tr>
<td>7</td>
<td>diamond</td>
<td>4</td>
<td>3</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>dioxide</td>
<td>153</td>
<td>16</td>
<td>0.27</td>
<td>0.32</td>
</tr>
<tr>
<td>9</td>
<td>fiber</td>
<td>77</td>
<td>21</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>10</td>
<td>film</td>
<td>94</td>
<td>52</td>
<td>0.87</td>
<td>0.73</td>
</tr>
<tr>
<td>11</td>
<td>gas</td>
<td>774</td>
<td>324</td>
<td>0.88</td>
<td>0.12</td>
</tr>
<tr>
<td>12</td>
<td>immune</td>
<td>2</td>
<td>2</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>13</td>
<td>inflammation</td>
<td>3</td>
<td>3</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>inhibitors</td>
<td>11</td>
<td>9</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>molecule</td>
<td>100</td>
<td>63</td>
<td>0.53</td>
<td>0.03</td>
</tr>
<tr>
<td>16</td>
<td>phenomenon</td>
<td>66</td>
<td>56</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>17</td>
<td>polymer</td>
<td>64</td>
<td>54</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>18</td>
<td>sensor</td>
<td>29</td>
<td>25</td>
<td>0.32</td>
<td>0.78</td>
</tr>
<tr>
<td>19</td>
<td>temperature</td>
<td>634</td>
<td>344</td>
<td>0.53</td>
<td>0.57</td>
</tr>
<tr>
<td>20</td>
<td>thermal</td>
<td>140</td>
<td>89</td>
<td>0.52</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 9.6: Search time (minutes) spent on each keyword (Group 1).

- **Number of items by SuperBook** means the total number of hits retrieved by SuperBook;
- **Number of items by E. SuperBook** means the total number retrieved by Enhanced SuperBook, i.e., the number of soft-link options.
- Because of the font problems in SuperBook, the soft-link hypertext is not available for carbon.
<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>No. of items by SuperBook</th>
<th>No. of items by E. SuperBook</th>
<th>SuperBook</th>
<th>E. SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A-SU</td>
<td>B-SU</td>
</tr>
<tr>
<td>1</td>
<td>absorption</td>
<td>159</td>
<td>84</td>
<td>0.45</td>
<td>0.68</td>
</tr>
<tr>
<td>2</td>
<td>amorphous</td>
<td>59</td>
<td>33</td>
<td>0.23</td>
<td>0.32</td>
</tr>
<tr>
<td>3</td>
<td>cohesive</td>
<td>1</td>
<td>1</td>
<td>0.28</td>
<td>0.43</td>
</tr>
<tr>
<td>4</td>
<td>density</td>
<td>194</td>
<td>95</td>
<td>0.22</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>distribution</td>
<td>573</td>
<td>309</td>
<td>0.10</td>
<td>0.67</td>
</tr>
<tr>
<td>6</td>
<td>elution</td>
<td>65</td>
<td>47</td>
<td>0.23</td>
<td>0.37</td>
</tr>
<tr>
<td>7</td>
<td>frequency</td>
<td>63</td>
<td>32</td>
<td>0.12</td>
<td>0.58</td>
</tr>
<tr>
<td>8</td>
<td>hydration</td>
<td>18</td>
<td>15</td>
<td>0.35</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>joint</td>
<td>15</td>
<td>10</td>
<td>0.17</td>
<td>0.55</td>
</tr>
<tr>
<td>10</td>
<td>methyl</td>
<td>94</td>
<td>52</td>
<td>0.55</td>
<td>0.82</td>
</tr>
<tr>
<td>11</td>
<td>nucleation</td>
<td>51</td>
<td>34</td>
<td>0.53</td>
<td>2.10</td>
</tr>
<tr>
<td>12</td>
<td>physical</td>
<td>180</td>
<td>122</td>
<td>0.65</td>
<td>0.88</td>
</tr>
<tr>
<td>13</td>
<td>radius</td>
<td>50</td>
<td>33</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>14</td>
<td>resin</td>
<td>58</td>
<td>26</td>
<td>0.77</td>
<td>0.42</td>
</tr>
<tr>
<td>15</td>
<td>rubber</td>
<td>33</td>
<td>25</td>
<td>1.90</td>
<td>1.53</td>
</tr>
<tr>
<td>16</td>
<td>scattering</td>
<td>22</td>
<td>12</td>
<td>0.27</td>
<td>0.55</td>
</tr>
<tr>
<td>17</td>
<td>temperature</td>
<td>634</td>
<td>344</td>
<td>0.22</td>
<td>0.67</td>
</tr>
<tr>
<td>18</td>
<td>tracer</td>
<td>100</td>
<td>71</td>
<td>0.08</td>
<td>0.70</td>
</tr>
<tr>
<td>19</td>
<td>uranium</td>
<td>6</td>
<td>5</td>
<td>0.28</td>
<td>0.55</td>
</tr>
<tr>
<td>20</td>
<td>vinyl</td>
<td>26</td>
<td>11</td>
<td>0.12</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 9.7: Search time (minutes) spent on each keyword (Group 2).
2. The Search Time;

The search time spent by each subject in the two treatment groups is listed in Tables 9.6 and 9.7. It can be seen that, for the IR tasks in which the exact keywords are available and only the location of the information is requested (From Submission of Queries to Retrieve Documents in Figure 9.2), the mean time spent by the subject using soft-link hypertext (Enhanced SuperBook) is about the same as that by Boolean search (SuperBook). This shows that, in the experimental environment, the subjects using the Enhanced SuperBook are not much affected by the introduction of the extra interface. They search and retrieve information almost as efficiently as the group using Boolean search.

Further analysis of the experimental results reveals the following facts:

- The soft-link hypertext treatment group spent roughly the same amount of time on each keyword. Its search time $T$ is not as much affected by the following factors as the other group:
  (a) The present position of the subject in the Information Space;
  (b) The number of hits included in the Information Space;
  (c) The location(s) of the information requested in the Information Space.

- The SuperBook group, on the contrary, has very good results if a negative, or only a few positive, search results are returned. In other words, the more hits found in Receive Search Results, the more difficult for the SuperBook users to spot any specific information in Retrieve Documents. This is a big problem for most of present large-scale information retrieval systems;

- The Boolean search is not as accurate as the soft-link hypertext in spotting detailed information. Sometimes, the subject in SuperBook treatment group finds that the Boolean search reveals positive results, but the information requested could not be found in the Information Space. For example, rubber compounds is requested in the experiment. Five hits are returned as search results by the Boolean search, though, in fact, the phrase is not included in the Information Space (the two words appear in the same paragraph for these five occasions). This is a common symptom in information retrieval, which makes information search and retrieval difficult for the user. The Enhanced SuperBook is much better on this point.
<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>No. of items by SuperBook</th>
<th>No. of items by E. SuperBook</th>
<th>SuperBook</th>
<th>E. SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>agent</td>
<td>43</td>
<td>21</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>biology</td>
<td>22</td>
<td>13</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>bonding</td>
<td>39</td>
<td>22</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>carbon</td>
<td>707</td>
<td>224</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>chemistry</td>
<td>452</td>
<td>158</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>condensation</td>
<td>70</td>
<td>54</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>diamond</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>dioxide</td>
<td>153</td>
<td>16</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>fiber</td>
<td>77</td>
<td>21</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>film</td>
<td>94</td>
<td>52</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>gas</td>
<td>774</td>
<td>324</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>immune</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>inflammation</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>inhibitors</td>
<td>11</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>molecule</td>
<td>100</td>
<td>63</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>phenomenon</td>
<td>66</td>
<td>56</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>polymer</td>
<td>64</td>
<td>54</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>sensor</td>
<td>29</td>
<td>25</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>temperature</td>
<td>634</td>
<td>344</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>thermal</td>
<td>140</td>
<td>89</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 9.8: User actions needed for search and retrieval (Group 1)
<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>No. of items by SuperBook</th>
<th>No. of items by E. SuperBook</th>
<th>SuperBook</th>
<th>E. SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A-SU</td>
<td>B-SU</td>
</tr>
<tr>
<td>1</td>
<td>absorption</td>
<td>159</td>
<td>84</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>amorphous</td>
<td>59</td>
<td>33</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>cohesive</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>density</td>
<td>194</td>
<td>95</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>distribution</td>
<td>573</td>
<td>309</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>elution</td>
<td>65</td>
<td>47</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>frequency</td>
<td>63</td>
<td>32</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>hydration</td>
<td>18</td>
<td>15</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>joint</td>
<td>15</td>
<td>10</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>methyl</td>
<td>94</td>
<td>52</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>nucleation</td>
<td>51</td>
<td>34</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>physical</td>
<td>180</td>
<td>122</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>radius</td>
<td>50</td>
<td>33</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>resin</td>
<td>58</td>
<td>26</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>rubber</td>
<td>33</td>
<td>25</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>scattering</td>
<td>22</td>
<td>12</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>temperature</td>
<td>634</td>
<td>344</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>tracer</td>
<td>100</td>
<td>71</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>uranium</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>vinyl</td>
<td>26</td>
<td>11</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 9.9: User actions needed for search and retrieval (Group 2)
3. The User Actions;

The numbers of user actions \( U \) needed by the subject in two treatment groups for searching and retrieving the requested information are listed in Tables 9.8 and 9.9. They can be analyzed as follows:

- It can be seen from the tables that the number of user actions needed by the Enhanced SuperBook group is much less than that of the SuperBook;
- One conclusion is that the soft-link hypertext improves greatly the user's global mobility in the Information Space. No matter where the subject's present position is, nor where the destination information is located, it usually needs only 2 – 3 user actions for the user to traverse from the present information node into another. From this aspect, the soft-link hypertext actually decreases the amount of cognitive overhead demanded from the user;
- The situation is very different with SuperBook:
  (a) If a Boolean search request can be totally rejected by the Superbooks (No hits can be found in the Information Space), this is very straightforward. Only a few user actions are needed to reach the conclusion;
  (b) If some search results are returned and the subject wants to locate the information requested via Search Forward, it is easy if the Boolean request is exclusively included (i.e., no search noise exists). However, it may turn out to be difficult if the user needs to go through a long list of noisy hits (produced by Boolean OR operator) before locating the actual information. This is one of the reasons why more user actions are needed for the SuperBook treatment group;
  (c) Alternatively, the subject can use Table of Content (TOC) to locate the information. Here, more user actions are usually needed to review the TOC and retrieve the related document before the locating process starts. Furthermore, the number of user actions needed is dependent on the physical location of the information inside the document;
  (d) The best way to browse by SuperBook lies on the combinational use of TOC and Search Forward (Always starting with TOC, and using Searching Forward once the document is retrieved). Unfortunately, this leads to a huge demand on the user's cognitive overhead and skills in using the browser.
4. Recall and precision;

Recall is the ratio of the relevant pieces of information retrieved by the system to the total number of relevant pieces of information in Information Space. According to the four subjects' qualitative judgements, the recall for both systems is very low, i.e., a lot of information, which is considered by the subjects as relevant, and therefore, should be included in the soft-link options, is not retrieved by either system.

The quantitative measurement of recall for the SuperBook and Enhanced SuperBook are listed in Tables 9.10 and 9.11. They are different from those in Table 9.1 as they are accessed here from the user's viewpoint. The more detailed analysis of the reasons is given in the next section (Section 9.5.4).

<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>No. of items</th>
<th>SuperBook</th>
<th>Enhanced SuperBook</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>agent</td>
<td>43</td>
<td>13.6</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>biology</td>
<td>22</td>
<td>48.9</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>bonding</td>
<td>39</td>
<td>72.2</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>carbon</td>
<td>707</td>
<td>33.8</td>
<td>224</td>
</tr>
<tr>
<td>5</td>
<td>chemistry</td>
<td>452</td>
<td>77.8</td>
<td>158</td>
</tr>
<tr>
<td>6</td>
<td>condensation</td>
<td>70</td>
<td>83.3</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>diamond</td>
<td>4</td>
<td>50.0</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>dioxide</td>
<td>153</td>
<td>73.9</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>fiber</td>
<td>77</td>
<td>53.8</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>film</td>
<td>94</td>
<td>25.5</td>
<td>52</td>
</tr>
<tr>
<td>11</td>
<td>gas</td>
<td>774</td>
<td>42.5</td>
<td>324</td>
</tr>
<tr>
<td>12</td>
<td>immune</td>
<td>2</td>
<td>100</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>inflammation</td>
<td>3</td>
<td>27.3</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>inhibitors</td>
<td>11</td>
<td>100</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>molecule</td>
<td>100</td>
<td>30.6</td>
<td>63</td>
</tr>
<tr>
<td>16</td>
<td>phenomenon</td>
<td>66</td>
<td>94.3</td>
<td>56</td>
</tr>
<tr>
<td>17</td>
<td>polymer</td>
<td>64</td>
<td>29.4</td>
<td>54</td>
</tr>
<tr>
<td>18</td>
<td>sensor</td>
<td>29</td>
<td>47.5</td>
<td>25</td>
</tr>
<tr>
<td>19</td>
<td>temperature</td>
<td>634</td>
<td>52.7</td>
<td>344</td>
</tr>
<tr>
<td>20</td>
<td>thermal</td>
<td>140</td>
<td>63.3</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 9.10: Recall & precision for Group 1.
Precision is defined as the ratio of the relevant information retrieved by the system to total number retrieved. As for the Enhanced SuperBook, the retrieved information is in the form of soft links, its precision should be the ratio of retrieved relevant soft-link options to the total number of options provided. Obviously, precision is a different measurement from that used in Experiment 1 (Section 9.4.2).

\[
P = \frac{\text{Number of soft links retrieved and relevant}}{\text{Total retrieved soft links}} \times 100 \%
\]

<table>
<thead>
<tr>
<th>No.</th>
<th>Keyword</th>
<th>SuperBook</th>
<th></th>
<th></th>
<th>Enhanced SuperBook</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. of items</td>
<td>Recall</td>
<td>Precision</td>
<td>No. of items</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>1</td>
<td>absorption</td>
<td>159</td>
<td>81.1</td>
<td>100</td>
<td>84</td>
<td>81.1</td>
<td>92.9</td>
</tr>
<tr>
<td>2</td>
<td>amorphous</td>
<td>.59</td>
<td>83.1</td>
<td>100</td>
<td>33</td>
<td>67.6</td>
<td>60.0</td>
</tr>
<tr>
<td>3</td>
<td>cohesive</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>density</td>
<td>194</td>
<td>70.8</td>
<td>100</td>
<td>95</td>
<td>62.4</td>
<td>78.9</td>
</tr>
<tr>
<td>5</td>
<td>distribution</td>
<td>573</td>
<td>60.6</td>
<td>100</td>
<td>309</td>
<td>54.5</td>
<td>95.8</td>
</tr>
<tr>
<td>6</td>
<td>elution</td>
<td>65</td>
<td>68.4</td>
<td>100</td>
<td>47</td>
<td>58.9</td>
<td>83.0</td>
</tr>
<tr>
<td>7</td>
<td>frequency</td>
<td>63</td>
<td>82.9</td>
<td>100</td>
<td>32</td>
<td>69.7</td>
<td>78.1</td>
</tr>
<tr>
<td>8</td>
<td>hydration</td>
<td>18</td>
<td>52.9</td>
<td>100</td>
<td>15</td>
<td>41.2</td>
<td>80.0</td>
</tr>
<tr>
<td>9</td>
<td>joint</td>
<td>15</td>
<td>88.2</td>
<td>100</td>
<td>10</td>
<td>88.2</td>
<td>80.0</td>
</tr>
<tr>
<td>10</td>
<td>methyl</td>
<td>94</td>
<td>8.0</td>
<td>100</td>
<td>52</td>
<td>7.9</td>
<td>94.2</td>
</tr>
<tr>
<td>11</td>
<td>nucleation</td>
<td>51</td>
<td>79.7</td>
<td>100</td>
<td>34</td>
<td>64.1</td>
<td>82.4</td>
</tr>
<tr>
<td>12</td>
<td>physical</td>
<td>180</td>
<td>75.9</td>
<td>100</td>
<td>122</td>
<td>64.1</td>
<td>32.8</td>
</tr>
<tr>
<td>13</td>
<td>radius</td>
<td>50</td>
<td>80.6</td>
<td>100</td>
<td>33</td>
<td>59.7</td>
<td>36.7</td>
</tr>
<tr>
<td>14</td>
<td>resin</td>
<td>58</td>
<td>52.7</td>
<td>100</td>
<td>26</td>
<td>47.3</td>
<td>69.2</td>
</tr>
<tr>
<td>15</td>
<td>rubber</td>
<td>33</td>
<td>44.0</td>
<td>100</td>
<td>25</td>
<td>42.7</td>
<td>84.0</td>
</tr>
<tr>
<td>16</td>
<td>scattering</td>
<td>22</td>
<td>91.7</td>
<td>100</td>
<td>12</td>
<td>91.7</td>
<td>83.3</td>
</tr>
<tr>
<td>17</td>
<td>temperature</td>
<td>634</td>
<td>52.7</td>
<td>100</td>
<td>344</td>
<td>23.4</td>
<td>48.3</td>
</tr>
<tr>
<td>18</td>
<td>tracer</td>
<td>100</td>
<td>53.2</td>
<td>100</td>
<td>71</td>
<td>47.9</td>
<td>81.7</td>
</tr>
<tr>
<td>19</td>
<td>uranium</td>
<td>6</td>
<td>100</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>80.0</td>
</tr>
<tr>
<td>20</td>
<td>vinyl</td>
<td>26</td>
<td>40.0</td>
<td>100</td>
<td>11</td>
<td>38.5</td>
<td>90.1</td>
</tr>
</tbody>
</table>

Table 9.11: Recall & precision for Group 2.
5. Response time: Considered as a defining element for hypertext in the literature [37], a short response time is very crucial and represents one of the most important requirements for any hypertext systems. The relative response time is considered here in the experiment, as the result of comparison between SuperBook and the Enhanced SuperBook. As these two systems share the same Document Space, the same hardware platform (Sun SPARC station) and the same programming language. Such a comparison should be meaningful.

It takes about 1 - 2 seconds for the SuperBook to respond, whereas for the Enhanced SuperBook, the time needed from the receipt of a hypertext request to the representation of local associations is about 1 - 4 seconds, depending on how many soft link options the conceptual index provides.

6. Presentation: Presentation refers to the form of display of the search results which influences the user's ability to utilize the retrieved materials.

The presentation of the Enhanced SuperBook is inferior to that of SuperBook. This is because an extra interface with another two levels of windows is introduced. Furthermore, problems like case-sensitivity, fonts, etc. are not considered. It is felt that although presentation of browsers is very important for any commercial IR systems, it is outside the scope of the present research, as long as it does not affect dramatically the user's behaviour during search and retrieval, nor influence substantially the outcome of the experiments.

9.5.4 Analysis of Results

1. The poor recall rate of both treatment groups is mainly caused by the following factors:

   • The Document Space used in both the SuperBook and the Enhanced SuperBook is not exhaustive. At the moment, only one year's collection (1991) of an ACS journal, Environmental Science and Technology, is used in both systems. The direct result is that a large number of soft-link options, which should be in the conceptual index, are not included just because they are not included at all in the Information Space;

   • The actual recall rates of global index used in both systems are lowerer than 100% (discussed in Experiment 1);
• The formulation of the conceptual index (soft links) is imprecise, based on a set of non-syntactic & non-semantic formulation rules used in the system.

2. As it is from the user's viewpoint, the precision for the Enhanced SuperBook (Tables 9.10 and 9.11) is different from that obtained in Experiment 1 (Tables 9.1 and 9.2). It provides a measurement of the proportion of soft links that is relevant and useful for the information user (Figures 9.4 and 9.5).

![Figure 9.4: Precision (Group 1) of the soft-link hypertext browser.](image)

3. Compared with that of the SuperBook, the response time for the soft-link hypertext browser should be acceptable. Usually, it takes about 1 - 4 seconds to respond to the information user's hypertext requests;

4. As the search formulation is conducted automatically once the hypertext mechanism is invoked [the user clicking word(s) in Page], the extra user effort needed to use soft-link hypertext is to check the conceptual index (soft links) and choose the ones in which the user is interested. Locally, this is the additional user overload brought about by the multiple options as the soft links. Nevertheless, it shows in the experiment that the addition is within a bearable range. In fact, it is in a different, even simpler form compared to those needed in other IR systems. This conclusion is based on the following observation:
Figure 9.5: Precision (Group 2) of the soft-link hypertext browser.

- Compared to the conventional hard-link hypertext, this checking and choosing process is an additional user cognitive overload brought by multiple run-time hypertext link options. However, a closer look reveals that this additional burden is within a reasonable range:

  (a) One extreme is that there are only a few soft-link options for the user, i.e., the retrieved soft-link list contains only few soft links [e.g., immune and inflammation in Group 1 (Table 9.4 and 9.5)]. In this case, the user effort and overload for using soft-link hypertext is exactly the same as that for a typed hard-link hypertext system;

  (b) With the increase of soft-link options, the user effort needed increases, as does the user overload. However, this increase is within a bearable range as checking the soft-link list is in fact a very simple and straightforward process. As each item in Glossary presents the syntax of a potential hypertext destination anchor, the choosing process is no more than the user's examining a list of hypertext destination anchors;

  (c) With the help of learning and intelligent mechanisms used in the soft-link hypertext model, this choosing process could be further shorten by the provision of short-list of popular links, as well as other user's
successful information trails. Further quantitative assessment of the intelligent mechanisms will be done in Experiment 3.

- Compared with that in the conventional IR systems, the user effort required here is in a different form. Taking a conventional IR system (SuperBook) as an example. There, the user effort in locating a specific information is represented by using features provided by the SuperBook (Search Forward, Searching Backward, etc.). Checking the Glossary and choosing a destination anchor in the soft-link hypertext browser is actually a different, even simpler procedure.

5. The user has fewer navigation worries with the Enhanced SuperBook. Even if he/she gets lost, less user effort is needed to become reoriented. A maximum of 2 - 3 user actions are needed to relocate a user to any terms, conceptual links, or information components, whereas reorientation in conventional IR systems like SuperBook might be a longer process;

6. The improved presentation of the Enhanced SuperBook (phrases in the Glossary) is much better than the original glossary words.

7. Finally, the soft-link hypertext browser does not affect the collection coverage of the information system.

9.5.5 Conclusion

In the experiment, the general acceptability of the soft-link hypertext browser is assessed, and the additional user cognitive overload brought by the integrated information model and the soft-link hypertext browser is evaluated at the qualitative level. The experiment shows that, although an extra cognitive burden has been brought to the user by making decisions among the multiple run-time soft link options, the actual addition is in a bearable range. In fact, it is in a different, even simpler form from some tools seen in conventional IR and conventional Hypertext systems (like typed hypertext, Fisheye, MAP, and Searching Forward, Searching Backward, etc.). Judged from the overall amount needed for locating a specific piece of information in the Information Space, the user effort required for the soft-link hypertext browser is actually less than those needed in other systems. The quantitative assessments are conducted in Experiment 3, where the amount of user effort and time needed for performing information-seeking tasks with both the Enhanced SuperBook and the SuperBook are compared.
9.6 Experiment 3 — Information Search/retrieval with Soft-link Hypertext

9.6.1 Purpose and Design

The purpose of Experiment 3 is to assess the effectiveness and efficiency of the soft-link hypertext system, as well as the intelligent mechanisms [user modelling, learning (LTM and STM) and reasoning] used in the system, by evaluating the user effort and time needed for performing some information retrieval tasks with the system. Its objects include the following:

1. To evaluate the overall performance of Enhanced SuperBook, as well as the user effort and time needed for some information-retrieval tasks:
   - In a free browsing environment;
   - In a task-driven environment.

2. To compare the Enhanced SuperBook with SuperBook with some IR tasks;

3. To assess the effectiveness of intelligent mechanisms used in the Enhanced SuperBook.

The hypothesis examined in the experiment is that, although an extra cognitive overload is introduced locally by using run-time soft-link hypertext, it could be well justified by the overall performance of the soft-link hypertext system, in the hypertext environment, over other conventional systems in the following three aspects:

1. Less user effort is needed in locating a specific piece of information;

2. Faster Information-seeking speed is reached by the provision of soft hypertext links, i.e., it takes less time for the user to search and retrieve some specific information;

3. Psychologically, the user feels more at ease during the information-seeking procedure and has fewer worries about disorientation, etc..

4. Finally, the intelligent mechanisms used in the Enhanced SuperBook could help to speed up this information-seeking process.
The design of the experiment was as follows:

1. Four treatments (independent variables) were used in the experiment:
   - Treatment 1: SuperBook;
   - Treatment 2: Enhanced SuperBook, with manually improved soft links;
   - Treatment 3: Enhanced SuperBook, with the original conceptual index and an implemented learning mechanism;
   - Treatment 4: Enhanced SuperBook, with the original conceptual index, but without learning mechanisms implemented.

2. 40 subjects were used in the experiment. They were assigned randomly to one of the treatment group, so that 4 x 10 subjects took part in the experiment;

3. Each subject was asked to look for some specific information. This included a free browsing task, and 4 information-retrieval and question-answering tasks. The instructions and questions for these tasks are listed in Appendix C;

4. The outcomes by these 4 x 10 subjects were collected and compared across the groups. Four dependent variables were used in the experiment. They included:
   - The time needed for completing a search and retrieval task;
   - The number of user actions needed for completing an IR task;
   - The accuracy rate of the information retrieved (the user's answers);
   - The user's satisfaction with the system.

9.6.2 Experimental Process

1. The experiment was conducted in the following steps:
   - The subject was assigned randomly to one of 4 treatment groups;
   - The subject was given a brief introduction to the system to be used, including its layout, features and functions;
   - The subject spent 10-minute on Task 1 (free browsing) to search, browse and read information according to his/her own interests and paces, with or without guidance;
   - This process investigates the outcomes in the subject's relaxed, informal and self-motivated information-retrieval procedure. It also provided each subject with an opportunity to get familiar with the system being used;
The number of pieces of information obtained during a 10-minute free browsing period, as well as the accuracy of the answers, was noted.

- The subject spent about 40 minutes on Task 2 (task-driven), to search and retrieve information for four different IR jobs, and answered questions based on the retrieved information.

This process investigates the outcomes of four treatment groups in a controlled information seeking environment. Four questions have been set in such a way that each of them represents a different IR task with its own emphasis. The user actions needed for finding the information to answer each question, as well as the time spent on each question, was noted.

2. On-line logging was installed in all systems to collect data automatically;

3. Think aloud was used during the experiment to record each subject’s psychological response during the information-seeking process.

9.6.3 Experimental Results & Analysis

The results of the experiment are summarized as the followings:

1. The outcomes of Task 1 for $4 \times 10$ subjects are listed in Tables 9.12 and 9.13. 10 questions are asked for Task 1, with the total score of 1. Each correct answer gets 0.1; wrong answers get 0; no answers are treated as wrong answers;

2. The results of Task 2 for the $4 \times 10$ subjects are summarized from Tables 9.14 to 9.21. The result and analysis are presented in the sequence of search jobs.

3. It is seen from tables and figures on the following pages that the experimental data representing each treatment group is not smooth, and is even quite rough. This is because, besides the influences of independent variables, the dependent variables are also affected by the following four characteristics of the subject:

   - Familiarity with computer hardware, especially the windows;
   - Domain knowledge and choices of keywords;
   - Reading speed;
   - Comprehension ability.

However, because of the random distribution of the subjects into each treatment group, the mean value of each group should still reflect influences by those independent variables exclusively.
## Experimental Results for Task 1 (Free Browsing)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>3.3</strong></td>
<td><strong>7.2</strong></td>
<td><strong>5.2</strong></td>
<td><strong>2.8</strong></td>
</tr>
</tbody>
</table>

Table 9.12: The number of pieces of information obtained by the user.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.6</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>0.8</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
<td>0.6</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>6</td>
<td>0.3</td>
<td>0.8</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>0.4</td>
<td>0.7</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td>0.4</td>
<td>0.8</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>9</td>
<td>0.2</td>
<td>0.7</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>0.3</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.3</strong></td>
<td><strong>0.68</strong></td>
<td><strong>0.5</strong></td>
<td><strong>0.21</strong></td>
</tr>
</tbody>
</table>

Table 9.13: The accuracy of answers for Task 1 (free browsing).
4. For Task 1, the idea is that no specific tasks were set for the subject, so that the information could be searched and retrieved in an informal and relaxed environment. Nevertheless, 10 questions were provided as guidance (Appendix C). Without these questions, most of the subject would not know what to do. It has been proven by the experiment that all subjects used these questions for their information retrieval;

5. The 10 questions used concern information from two articles in Environmental Science and Technology, 1991. The two articles were written by the same authors, their contents overlap though they are basically on different topics. The questions have been arranged in such a way that cross-hypertext-links are needed to read efficiently. This provides an ideal hypertext applicational environment, which can be characterized by the following factors:

- The information requested is at the sentence level;
- The information requested is distributed in the Information Space;
- The location of requested information in the database is not important;
- The movement among the information nodes (to and from) needs to be fast and convenient.

6. Subjects in Group 1 (SuperBook) have two approaches to obtain information:

- Look up — Table of Content (TOC) — Page — Information;
- Look up — Search Forward — Information.

Subjects in Group 2, 3 and 4 (the Enhanced SuperBook) use the soft-link hypertext approach for the problem, namely:

- Hypertext — Glossary (soft links) — Information.

As the Enhanced SuperBook provides a more powerful hypertext tool than Search Forward in SuperBook, the result of the experiment is expected to favour the former over the latter, in the hypertext applicational environment;

7. Table 9.12 shows the number of pieces of information each subject in these four treatment groups obtained during the 10-minute free browsing. It can be seen that the subject using Treatment 2 retrieved more information (average 7.2) than the other groups. This shows that the subjects using the Enhanced SuperBook with the improved soft links could search and retrieve faster (more efficiently) than the other treatment groups. This can also be seen from Figure 9.6.
8. Also as shown in Figure 9.6, an interesting fact about Treatment 3 is that it starts at about the same level as that of Treatment 4, but rises so that it eventually approaches that of Treatment 2. This proves the point that the soft-link hypertext model with a learning mechanism could improve the quality of its links automatically and help in making the information retrieval more efficient. There exists a turning point at which the effective learning happens, i.e., the useful soft links start becoming dominate (Subject 3 in this case). All subjects after the turning point behave better. This proves the proposition of the Kohonen learning rule, i.e. once a Kohonen layer is correctly trained, it remains so for all $t$ (Section 6.8);

![Figure 9.6: The information obtained in Task 1](image)

Nevertheless, it can also be seen from the experimental data that the appearance of the turning point is somewhat arbitrary;

9. It can be seen in Tables 9.12 and 9.13 (not so obvious in Figure 9.6) that the outcome of Treatment 4 is not as good as that of Treatment 1. This reveals the fact that the soft-link hypertext system without a learning mechanism does not work as efficient as SuperBook. This is because the subject using Treatment 4 needs to go through the whole soft-link list every time, which only provides a slightly better service than *Search Forward*, but not as good as the latter supplemented by *Table of Content (TOC)* in SuperBook;
10. Table 9.13 and Figure 9.7 show the accuracy rates of the retrieved information by these four treatment groups:

- With the soft links pointing the information location to the sentence level, it is not surprising that the accuracy rate by Treatment 2 (the Enhanced SuperBook with improved soft links) is the highest;
- SuperBook provides location pointers at two different levels: Search Forward at the paragraph level, whereas Table of Content at the document level. Unfortunately, Search Forward does not work very well in the present version of SuperBook. The subject has to depend on TOC to obtain information. It, therefore, becomes a difficult and time-consuming process to move between the two articles via TOC. This is why the accuracy rate by Treatment 1 (SuperBook) is quite low;
- The accuracy rates by Treatment 4 is low, mainly because reviewing the whole list of soft link options slows down the speed of information retrieval;
- Treatment 3 improves this when the system starts learning. It can be seen that the accuracy rates of answers increase from Subject 3 in the group, which confirms with that in Figure 9.6. The average accuracy rate of Treatment 3 is 0.5 (Table 9.13).
### Experimental Results for Question 1 (Task 2)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>4</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>3</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>3</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Average</td>
<td>9</td>
<td>2.4</td>
<td>5.2</td>
<td>11.9</td>
</tr>
</tbody>
</table>

Table 9.14: Number of User actions needed for Question 1.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.97</td>
<td>2.18</td>
<td>8.75</td>
<td>9.02</td>
</tr>
<tr>
<td>2</td>
<td>2.30</td>
<td>1.57</td>
<td>7.03</td>
<td>8.32</td>
</tr>
<tr>
<td>3</td>
<td>8.92</td>
<td>2.05</td>
<td>6.58</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>6.55</td>
<td>1.32</td>
<td>6.79</td>
<td>6.97</td>
</tr>
<tr>
<td>5</td>
<td>7.03</td>
<td>5.57</td>
<td>3.48</td>
<td>5.35</td>
</tr>
<tr>
<td>6</td>
<td>9.87</td>
<td>2.48</td>
<td>1.35</td>
<td>3.83</td>
</tr>
<tr>
<td>7</td>
<td>5.80</td>
<td>1.32</td>
<td>1.07</td>
<td>8.00</td>
</tr>
<tr>
<td>8</td>
<td>7.79</td>
<td>1.20</td>
<td>3.63</td>
<td>8.87</td>
</tr>
<tr>
<td>9</td>
<td>10.0</td>
<td>0.97</td>
<td>1.95</td>
<td>7.04</td>
</tr>
<tr>
<td>10</td>
<td>8.32</td>
<td>1.39</td>
<td>2.44</td>
<td>8.77</td>
</tr>
<tr>
<td>Average</td>
<td>7.35</td>
<td>2.00</td>
<td>4.31</td>
<td>7.62</td>
</tr>
</tbody>
</table>

Table 9.15: Time (minutes) spent on Question 1
11. The answer for Question 1 is in the article "Sources of Fine Organic Aerosol. 1. Charbroilers and Meat Cooking Operations", (Environment Science and Technology, pp. 1112-1125, 1991). It includes 13.5 A4 printed pages, or 32 screen pages on SuperBook. The location of the required information in SuperBook can be roughly described as the following:

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Features</th>
<th>Hits</th>
<th>User actions/readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(any)</td>
<td>TOC/Next Page</td>
<td>1</td>
<td>32/difficult</td>
</tr>
<tr>
<td>cooking</td>
<td>Lookup/Search Forw</td>
<td>51</td>
<td>51/difficult</td>
</tr>
<tr>
<td>cooking process</td>
<td>Lookup/Search Forw</td>
<td>6</td>
<td>6/difficult</td>
</tr>
<tr>
<td>hamburger</td>
<td>Lookup/Search Forw</td>
<td>31</td>
<td>15/easy</td>
</tr>
<tr>
<td>hamburger surface</td>
<td>Lookup/Search Forw</td>
<td>3</td>
<td>7/easy</td>
</tr>
<tr>
<td>hamburger outer surface</td>
<td>Lookup/Search Forw</td>
<td>1</td>
<td>5/easy</td>
</tr>
</tbody>
</table>

As this article is the exclusive one on the topic in the whole database, it is easy to be identified with any of the keywords hinted in the question (e.g., substances, outer surface, hamburger, cooking process). However, finding the information after locating the document is quite difficult because:

- The article is quite long (32 screens);
- The information requested is located in the later half of the article.

Using *Search Forward* could make the search and retrieval quite straightforward, especially when the Boolean search is adopted.

12. Treatment 2 presents the best result in this case, with average number of user actions of 2.4 (Table 9.14). It presents a quite smooth curve in Figure 9.8, which shows that information retrieval for the subject in this group is quite easy;

13. Ideally, if the Boolean search is used for *Search Forward*, the number of user actions needed by the subject using Treatment 1 for information for Question 1 should be roughly the same as that in soft-link hypertext. However, from the experimental result (Table 9.14 and Figure 9.8), it can be seen that more actions are actually used in Treatment 1 than in Treatment 2.
One reason might be that the Boolean search is still too difficult for most of the subjects, as reported in other IR studies. Another reason is that although using the Boolean search could help in locating quickly the node in which the information is included, it quite often results in too many highlighted terms (keywords) on one screen for the subject to concentrate on the reading and obtaining any valuable information. This leads to an increase in the total time needed for information retrieval. What makes it worse is that the more keywords a subject uses in the Boolean search, the better the search result should approach that of the soft-link hypertext; unfortunately, the problem of readability becomes more serious (Figure 9.9). Searching via the soft-link hypertext system does not have this problem.

14. The curve of Treatment 3 in Figure 9.8 also shows the fact that the system has been trained (from Subject 4 in this case). As the result, the number of user actions and the time needed for later subjects decrease to about the same level as that of Treatment 2;

15. In this case, the soft link anchor appears as the last entry on the list. This probably explains why it takes longer (4 subjects) for the system to be trained. Some of the earlier subjects might have given up before they reach the proper soft-link option;
Air pollution damage to museum collections has emerged as a major issue in art conservation. Outdoor pollution and more specifically the widespread damage caused by acid deposition on stone, has been recognized as a threat to cultural property for a number of years (1-3). Damage to works of art due to exposure to indoor pollutants has recently received increasing attention (4, 5). Recent studies have documented the presence of ozone in concentrations ranging from 3 to 80% of the corresponding outdoor levels (6-7). Other factors contributing to the deterioration of cultural property include: excessive humidity, high salinity and solar radiation, and water contamination. Among other substances, the pollutants that are potentially damaging to museum collections, especially nitric oxide, sulfur dioxide, and formaldehyde, have been the subject of recent surveys. With respect to other pollutants, little is known regarding their levels in indoor settings and the corresponding potential for damaging effects.

Accordingly, we have undertaken a survey of selected air pollutants at one southern California museum. The selection of museum involved three major: diversity of the collections, outdoor/indoor exchange characteristics, and geographical location. The Los Angeles area, where outdoor pollution concentrations exhibit pronounced thermal, seasonal, and spatial variations. The five museums surveyed included several major art museums with many types of collections, a natural history museum, an archeological museum, an ethnology museum, one of the largest libraries in the United States, and several historical buildings.
16. The time spent by subjects in the four treatment groups (Figure 9.10) represents roughly the same picture, though it is more affected by other factors like the subjects' reading and comprehension ability. The subjects using Treatment 2 spent the least time in searching and retrieving information; The time spent by the subjects in Treatment 3 decreases from Subject 4 (the turning point). Whereas the subject using Treatment 4 needs more actions in obtaining the information than those in Treatment 1, the time spent by those two groups are almost the same. This is because the user generally moves faster with the soft-link hypertext system;

17. This experiment reveals a very interesting fact about the Boolean search and its relationship with the soft-link hypertext. The purpose of Boolean search and the soft-link hypertext are almost the same — to provide more information about the destination node. However, it seems from the result of the experiment that the former is still too difficult for an ordinary user, and it is further limited by its readability problems (Point 13). It is also strongly affected by the user's ability in choosing proper keywords. The last point is further shown and investigated in the later part of the experiment;
Experimental Results for Question 2 (Task 2)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Average</td>
<td>16.6</td>
<td>3.9</td>
<td>3.9</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Table 9.16: Number of User actions needed for Question 2.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.32</td>
<td>0.84</td>
<td>1.54</td>
<td>1.06</td>
</tr>
<tr>
<td>2</td>
<td>10.0</td>
<td>1.21</td>
<td>1.09</td>
<td>1.28</td>
</tr>
<tr>
<td>3</td>
<td>8.93</td>
<td>1.32</td>
<td>0.54</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>5.45</td>
<td>1.08</td>
<td>2.92</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>9.04</td>
<td>1.97</td>
<td>0.78</td>
<td>1.32</td>
</tr>
<tr>
<td>6</td>
<td>10.0</td>
<td>0.95</td>
<td>1.12</td>
<td>1.17</td>
</tr>
<tr>
<td>7</td>
<td>5.21</td>
<td>1.05</td>
<td>1.20</td>
<td>1.02</td>
</tr>
<tr>
<td>8</td>
<td>7.88</td>
<td>1.41</td>
<td>1.37</td>
<td>1.14</td>
</tr>
<tr>
<td>9</td>
<td>4.98</td>
<td>1.25</td>
<td>1.08</td>
<td>1.33</td>
</tr>
<tr>
<td>10</td>
<td>6.73</td>
<td>1.16</td>
<td>1.15</td>
<td>1.85</td>
</tr>
<tr>
<td>Average</td>
<td>7.65</td>
<td>1.22</td>
<td>1.23</td>
<td>1.28</td>
</tr>
</tbody>
</table>

Table 9.17: Time (minutes) spent on Question 2.

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Features</th>
<th>Hits</th>
<th>User actions/readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>air pollution</td>
<td>TOC/Next Page</td>
<td>135</td>
<td>Impossible</td>
</tr>
<tr>
<td>air pollution museums</td>
<td>TOC/Next Page</td>
<td>13</td>
<td>15/easy</td>
</tr>
<tr>
<td>museums</td>
<td>Lookup/Search Forw</td>
<td>17</td>
<td>11/easy</td>
</tr>
<tr>
<td>air pollution</td>
<td>Lookup/Search Forw</td>
<td>135</td>
<td>Impossible</td>
</tr>
<tr>
<td>air pollution museums</td>
<td>Lookup/Search Forw</td>
<td>13</td>
<td>Impossible</td>
</tr>
</tbody>
</table>

19. The information is located on the third screen of the article (Introduction), which should be reasonably easy to identify.

Figure 9.11: Number of User actions needed for Question 2

Here, the question is set in such a way that only inefficient keywords are used (except for museum), which means that some words used in the question are too popular to be good candidates for effective Boolean search. Proper use of Boolean operators on multiple keywords is even more difficult. In this case, many features of SuperBook must be used altogether to retrieve the information efficiently;
20. The subjects in Treatment 1 need more user actions than other groups to retrieve the information;

21. The number of user actions needed for the other three treatment groups are about the same, and the effect of learning is not obvious in this case. This is because only 15 soft-link options are related to museums, which takes only one screen. Here, learning could not dramatically decrease the number of user actions and the time needed for locating the appropriate information.

![Graph of Time Used vs. Treatment](image)

**Figure 9.12: Time spent on Question 2**

22. One important fact revealed from the experiment is that TOC in SuperBook is not of much help in pinpointing detailed information. Although the combined use of the Boolean search and TOC may speed up the locating process, it makes reading more difficult, unless the wanted information itself is highlighted; otherwise, it is very easy for the user to miss the real useful information;

23. *Search Forward* in SuperBook is also not good enough for pinpointing the detailed information. Quite often, the user needs to go through a long list of hits before locating any useful information. More auxiliary tools are needed to provide further help.
Experimental Results for Question 3 (Task 2)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>3</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>4</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>5</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>6</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
<td>4</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>Average</td>
<td>11.6</td>
<td>4.3</td>
<td>8.8</td>
<td>12.8</td>
</tr>
</tbody>
</table>

Table 9.18: Number of User actions needed for Question 3.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.04</td>
<td>1.27</td>
<td>10.0</td>
<td>7.04</td>
</tr>
<tr>
<td>2</td>
<td>6.33</td>
<td>2.40</td>
<td>3.35</td>
<td>3.73</td>
</tr>
<tr>
<td>3</td>
<td>6.01</td>
<td>0.93</td>
<td>1.41</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>5.54</td>
<td>1.12</td>
<td>1.83</td>
<td>7.68</td>
</tr>
<tr>
<td>5</td>
<td>0.62</td>
<td>1.83</td>
<td>2.63</td>
<td>1.21</td>
</tr>
<tr>
<td>6</td>
<td>10.0</td>
<td>1.44</td>
<td>9.85</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>1.87</td>
<td>0.96</td>
<td>6.54</td>
</tr>
<tr>
<td>8</td>
<td>3.93</td>
<td>2.03</td>
<td>1.21</td>
<td>10.0</td>
</tr>
<tr>
<td>9</td>
<td>4.45</td>
<td>1.53</td>
<td>2.32</td>
<td>6.80</td>
</tr>
<tr>
<td>10</td>
<td>8.78</td>
<td>0.51</td>
<td>1.93</td>
<td>8.33</td>
</tr>
<tr>
<td>Average</td>
<td>4.95</td>
<td>1.20</td>
<td>3.55</td>
<td>6.22</td>
</tr>
</tbody>
</table>

Table 9.19: Time (minutes) spent on Question 3.

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Features</th>
<th>Hits</th>
<th>User actions/readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>methane</td>
<td>TOC/Next Page</td>
<td>107</td>
<td>Impossible</td>
</tr>
<tr>
<td>methane emission</td>
<td>TOC/Next Page</td>
<td>6</td>
<td>10/easy</td>
</tr>
<tr>
<td>rice</td>
<td>TOC/Next Page</td>
<td>36</td>
<td>10+/easy</td>
</tr>
<tr>
<td>rice fields</td>
<td>TOC/Next Page</td>
<td>16</td>
<td>10/easy</td>
</tr>
<tr>
<td>methane</td>
<td>Search Forw</td>
<td>107</td>
<td>Impossible</td>
</tr>
<tr>
<td>methane emission</td>
<td>Search Forw</td>
<td>6</td>
<td>5/difficult</td>
</tr>
<tr>
<td>rice</td>
<td>Search Forw</td>
<td>36</td>
<td>16/easy</td>
</tr>
<tr>
<td>rice fields</td>
<td>Search Forw</td>
<td>16</td>
<td>4/easy</td>
</tr>
</tbody>
</table>

25. The answer for Question 3 is on the first screen page of the article. This means that, once the subject locates the article, the information requested is found. Meanwhile, as the number of hits for each of these keyword(s) is not very high, the subject using Treatment 1 is expected to behave well in this case;

![Figure 9.13: Number of User actions needed for Question 3](image)

26. Question 3 is asked in a longer context, so that it is more difficult for the subject to identify the effective keyword(s) used for information retrieval. As shown
in Figure 9.13, the outcome from the four treatment group represent much rough curves. Still, Treatment 2 represents the best picture, with the minimum average number of user actions for answering the question (4.3). Treatment 1 and 4 show some irregular subjects’ behaviour during the information retrieval. In Treatment 3, the effect of learning is very obvious (except for Subject 6). Because the list of soft links is quite long in this case, the number of user actions and time spent on the task could be reduced greatly once the system is trained;

Figure 9.14: Time spent on Question 3

27. Again, in Figure 9.14, the time spent on searching and retrieving information to answer Question 3 by each treatment group reveals a similar picture: the subject in Treatment 1 could have very good results. But, it is up to each subject’s skills of identifying proper keywords, using the Boolean search, and choosing the suitable features of SuperBook, whereas that for Treatment 2 is more consistent and unaffected either by the location of the requested information, nor the user’s present position in the Information Space;

28. The average time spent by subjects in Treatment 3 is shorter (3.55 minute) than those in Treatments 1 and 4. This shows that with learning mechanism, the soft links can be improved and become more helpful in information retrieval;
### Experimental Results for Question 4 (Task 2)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>10</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>17</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>15</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>25</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>8</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>7</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>7</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>12</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>12.6</strong></td>
<td><strong>12.2</strong></td>
<td><strong>14.2</strong></td>
<td><strong>12.5</strong></td>
</tr>
</tbody>
</table>

Table 9.20: Number of User actions needed for Question 4.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Treatment 1</th>
<th>Treatment 2</th>
<th>Treatment 3</th>
<th>Treatment 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.60</td>
<td>3.43</td>
<td>5.44</td>
<td>9.57</td>
</tr>
<tr>
<td>2</td>
<td>6.80</td>
<td>4.07</td>
<td>10.0</td>
<td>1.32</td>
</tr>
<tr>
<td>3</td>
<td>7.03</td>
<td>7.20</td>
<td>6.00</td>
<td>6.37</td>
</tr>
<tr>
<td>4</td>
<td>1.17</td>
<td>5.51</td>
<td>5.94</td>
<td>2.21</td>
</tr>
<tr>
<td>5</td>
<td>5.00</td>
<td>4.04</td>
<td>1.59</td>
<td>5.83</td>
</tr>
<tr>
<td>6</td>
<td>5.49</td>
<td>10.0</td>
<td>2.33</td>
<td>3.02</td>
</tr>
<tr>
<td>7</td>
<td>1.25</td>
<td>3.93</td>
<td>7.16</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>10.0</td>
<td>3.43</td>
<td>2.51</td>
<td>7.49</td>
</tr>
<tr>
<td>9</td>
<td>5.44</td>
<td>3.97</td>
<td>5.77</td>
<td>2.94</td>
</tr>
<tr>
<td>10</td>
<td>3.71</td>
<td>5.31</td>
<td>7.03</td>
<td>5.48</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>4.85</strong></td>
<td><strong>5.09</strong></td>
<td><strong>5.34</strong></td>
<td><strong>5.42</strong></td>
</tr>
</tbody>
</table>

Table 9.21: Time (minutes) spent on Question 4.
29. Question 4 is set on the article "Indoor Air Pollutants from Unvented Kerosene Heater Emissions in Mobile Homes..." (Environmental Science and Technology, pp. 1732 - 1738, 1991). It includes 7 A4 printed pages, or 21 screen pages on SuperBook;

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Features</th>
<th>Hits</th>
<th>User actions/readability</th>
</tr>
</thead>
<tbody>
<tr>
<td>kerosene</td>
<td>TOC/Next Page</td>
<td>35</td>
<td>7/easy</td>
</tr>
<tr>
<td>kerosene heaters</td>
<td>TOC/Next Page</td>
<td>25</td>
<td>7/difficult</td>
</tr>
<tr>
<td>pollutants</td>
<td>TOC/Next Page</td>
<td>284</td>
<td>Very difficult</td>
</tr>
<tr>
<td>air pollutants</td>
<td>TOC/Next Page</td>
<td>112</td>
<td>Difficult</td>
</tr>
<tr>
<td>kerosene</td>
<td>Search Forw</td>
<td>35</td>
<td>10/easy</td>
</tr>
<tr>
<td>kerosene heaters</td>
<td>Search Forw</td>
<td>25</td>
<td>15/difficult</td>
</tr>
<tr>
<td>pollutants</td>
<td>Search Forw</td>
<td>284</td>
<td>Very difficult</td>
</tr>
<tr>
<td>air pollutants</td>
<td>Search Forw</td>
<td>112</td>
<td>Very difficult</td>
</tr>
</tbody>
</table>

Figure 9.15: Number of User actions needed for Question 4

30. The information is located on the first screen page of the article. Again, once the subject obtains the article, the required information is located;
31. The question is asked in a very long sentence so that choosing proper keywords for information retrieval is very difficult. One of the option “air pollutants” is misleading, though it appears in the original sentence in the article. As it is too popular in the Information Space (inefficient keywords), it provides no help in identifying the useful information. The best way for Question 4 is to use TOC to locate the article. Once the article is found, the information can be identified as it is on the first page of the article;

32. Also because of the way the question is set (no proper source anchor is hinted in the question), the use of soft-link hypertext is almost impossible in this case (i.e., no effective soft hypertext links are provided);

33. In this case, the number of user actions and the time needed for information retrieval depend totally on the subject’s choices of keywords and combined use of different features. It shows that when no proper soft links are provided, Treatments 2, 3 and 4 behave roughly the same as that of Treatment 1.
9.6.4 Conclusion

Experiment 3 investigates the soft-link hypertext system (the Enhanced SuperBook) in real IR environments, and compares its behaviour with the SuperBook. The purpose is to assess the system, as well as the learning mechanism used in it, in a specific IR environment — hypertext. In this dissertation, such an environment is characterized as the following:

- Search formulation is not a problem for the information user;
- The information sought (the destination anchor) is located at the sentence level;
- The initial location (of the source anchor, or the user in the Information Space) would not affect the IR result and efficiency;
- The content of the destination information node is more important than other attributes (e.g., author names, the date of publication, etc.);
- Movement between the information nodes need to be fast and convenient.

The following points can be concluded from the experiment:

1. Without proper soft hypertext links provided, the Enhanced SuperBook acts about the same as SuperBook in the environment discussed (Points 32 and 33);
2. The Enhanced SuperBook shows a great improvement over the SuperBook in the hypertext environment, when soft hypertext links are provided. It provides much more consistent and effective IR services for the user (Treatment Group 2 vs Treatment Group 1, presented in Points 7, 13, 14, 20, 21 and 27);
3. It is very important to use learning mechanisms to improve the quality of soft links, which could further speed up the information search and retrieval (Comparison among Treatment Groups 2, 3 and 4, as presented in Points 8, 9, 14);
4. The more items are included in a conceptual index (soft links), the more difficult it is to learn. However, once the system starts learning, the more improvements it brings to information retrieval. In other words, if only a few items are included, the effect of learning is not obvious at all (Points 14 and 15).

The conclusion of Experiment 3 is that, in the hypertext environment tested, the Enhanced SuperBook presents a much better IR system for information search and retrieval. It facilitates a big improvement in both speed and quality of the IR services it provides for the information user.
9.7 Experiment 4 — An Intelligent Information System

9.7.1 Purpose and Design

Experiment 4 was originally designed to assess the roles of different IR models and compare the soft-link hypertext browser with others in an integrated information retrieval environment.

As discussed in Section 8.4, each IR model and system has its own strengths and weaknesses. Nevertheless, the soft-link hypertext system combine different models, including the Boolean search, hard-link hypertext and soft-link hypertext, in one infrastructure; it should be able to provide an integrated information service for different levels of information needs.

The design of the experiment is described as follows:

1. Three treatments were to be used in the experiment: *SuperBook* for the Boolean search model, *the Enhanced SuperBook with hard-link hypertext* (Figure 8.12) for the conventional hypertext model and *the Enhanced SuperBook with soft-link hypertext browser* (Figure 8.14) for the soft-link hypertext model;

2. The subject was to be assigned to one of the three treatment groups and was to be asked to fulfill one of the following three categories of information retrieval tasks:

   • **Information Introduction.** Each subject in this group would not be given any background information before the session. The subject would search and browse for information totally on his/her own interests and answer some questions at the end of the session;

   • **Information Expansion.** Each subject in this group would be given some background information by reading a short paragraph about the domain before starting search and retrieval, and would answer questions according to the information retrieved during the session;

   • **Information Pinpointing.** For this group, each subject would need to read several paragraphs and obtain some quite detailed information before starting information search and retrieval, and finally would answer questions based on the information retrieved.
3. The questions asked would reflect the related information needs for each of the three groups;

4. Eventually, the outcomes from these nine groups would be summarized in the format of the following table and compared across the group, before some conclusions could be drawn.

<table>
<thead>
<tr>
<th>Application Environment</th>
<th>Conventional IR</th>
<th>Hard-link Hypertext</th>
<th>Soft-link Hypertext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Introductory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Expansion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Pinpointing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9.7.2 Difficulties in Running Experiment 4

It was not realized, until starting to set out the detailed steps and implementation of the experiment, that running Experiment 4 in a small-scale controlled environment was virtually impossible. As a result, Experiment 4 was stopped. The difficulties that forced the experiment to be cancelled were as follows:

1. Building up a hard-linked hypertext network within the Ph.D period would be extremely difficult. Such a task was demanding in two aspects: Firstly, I needed to be quite knowledgeable and very well informed about the domain (chemistry) before I could be able to link any information components; secondly, it would be very time-consuming to construct a reasonably-sized hypertext net, whilst a unduly small-scale net was worthless both from the experiment's point of view, and for the interests of the subject;

2. It would be very difficult to measure the results of information retrieval in such a controlled experimental environment. For example, for the group of Information Introduction, it would be very difficult to choose an area, which was new and unknown to all the subjects. Once a subject knew something about the chosen topic, the outcome of such a experiment would be changed into Information Expansion, rather than reflecting accurately that of Information Introduction. Other problems included how the questions should be asked to assess the quality of the information retrieval, whether these questions should be shown to the subject and when. On the one hand, revealing questions too early would give hints about the domain knowledge and keywords; on the other
hand, without showing the subjects the questions before (or during) the ses-
session, what would the subject be expected to do in the controlled laboratory
environment?

9.7.3 The Impact on Information Retrieval

Although it is unfortunate that I was not able to run Experiment 4 because of the
difficulties listed above, the idea for such an experiment is still worth pursuing. Whilst
these difficulties provide evidence that a hypertext model like the one proposed in
this dissertation is urgently needed in the community, running such an experiment
could constitute further support for the model, which remains an open question
(Section 10.3.2). Should a larger soft-link hypertext system be set up publicly, and
operated continuously for a period of time (e.g., 6–12 months), a group of useful
soft links could become prominent. This group could either become permanent links
(hard links), or provide a basis from which hard links could be set up by domain
experts. Then, running Experiment 4 in a natural environment, with real information
users who access the system with real information needs, should produce some very
interesting results for information retrieval.

9.8 Conclusion

Although the soft-link hypertext model has been assessed theoretically in Chapter 8,
and its strengths shown in accommodating an integrated IR environment in which
faster and more convenient IR services could be provided for different level of infor-
mation retrieval, the model and the implemented system (the Enhanced SuperBook)
need to be further evaluated in some practical operations. Therefore, the following
four experiments have been designed in my research to assess various parts of the
model, to test how good they are and to find out the reasons.

1. Experiment 1 tests the automatic index formulation mechanism. It shows that a
reasonably good conceptual index structure could be formulated automatically,
within an acceptable cost of time and memory;

2. Experiment 2 confirms that such a data structure, as well as the soft-link hy-
pertext browser, is acceptable for the user. Some cognitive overload is brought
about by the introduction of the additional features; nevertheless, they are fully justified by the overall performance of the system;

3. Experiment 3 tests the overall performance of the soft-link hypertext system. Its results show that, with the soft-link hypertext system, the user searches and retrieves information faster and easier. Moreover, the learning mechanism could further speed up this process;

4. Finally, Experiment 4 would have tested the soft-link hypertext model in an integrated IR environment. It was given up because of the limitations of time and experimental conditions in the present research. However, it could well be possible for a larger trial system and could produce some very interesting results.
Chapter 10
Conclusion

Chapter 10 summarizes the validity of the original hypotheses and draws relevant conclusions showing how the remit of the dissertation has been met. This evaluation is based on lessons learnt from the development of the soft-link hypertext model and the experimental results obtained from the assessment of an intelligent hypertext system, the Enhanced Super-Book. Whilst the soft-link hypertext model constitutes the main contribution of this research, its position in the present IR community, the limitations of the model and its impact on future development of information systems are also discussed here.

10.1 Summary of the Work

The essence of the work presented in this dissertation is to study the possibility of constructing automatically a data structure which is suitable for information retrieval (especially hypertext), using statistical information to further improve and adjust such a structure, and applying it effectively in information search and retrieval. There are the following three different aspects of work involved here:

1. The design of such a data structure;
2. The development of a model which supports such a data structure, and more importantly, facilitates the mechanisms for automatic generation, improvement and applications of the structure;
3. The adoption of the model in information retrieval to improve the efficiency and effectiveness of the present IR systems.

As a result, the soft-link hypertext model has been developed and implemented (Chapter 5, 6, 7 and 8). Relying on a more sophisticated data structure called the conceptual index, the model shares a lot of similarities with most of the existing hypertext systems, but differs on its interpretation of hypertext elements (node, anchor and link). Therefore, the soft-link hypertext model is blessed uniquely with auto-generation, self-adjustment and versatility; it is more suitable for large-scale, user-centered and intelligent information retrieval environment.
With such a model, it also becomes possible to prove the hypothesis examined in this dissertation. This can be shown as follows:

1. A conceptual index has been generated automatically by the conceptual index formulation mechanism employed in the model (Chapters 5 and 9);
2. The conceptual index facilitates both the global and local access. It helps in improving the effectiveness and efficiency of IR systems (Chapters 5, 8 and 9);
3. Learning via statistical information has been applied in the model to tune the data structure and accumulate its past experiences (Chapters 6, 7 and 9);
4. The machine intelligence has been adopted to provide a versatile, flexible and user-centered system (Chapters 6, 7, 8 and 9).

10.2 Contribution to IR

In assessing the relevance and contribution of this work to IR, it is first necessary to assess where it stands in relation to the previous work. By comparing it with other projects, my contribution to the development of information retrieval and machine learning, as well as the limitation of the methodologies adopted, can be identified.

10.2.1 The Previous Work

The most closely related work is Savoy's Bayesian network [139], Frisse's belief network [68], Turtle & Croft's inference networks [159] and their applications in IR. Those approaches and the soft-link hypertext model all adopt the idea of probability in information retrieval and use a kind of activation spreading on the semantic network installed on the global index as an inferencing procedure. However, my model is different in the nature of the semantic network used (which is more suitable for hypertext), in the automatic generation of the network and in attaching a weight to each connections of the semantic network so that its structure could be self-adjusted via statistical information. In this way, not only does the hypertext structure become dynamic and evolvable, but more importantly, some extra auxiliary information is available for the information user.

Other research on the applications of the connectionist network in IR [8] [89] are related to the soft-link hypertext model in the way that both forward activation
spreading and backward learning are utilized during information retrieval. Whilst the previous work is only applicable for bibliographical IR tasks, the soft-link hypertext model is targeting at the full-text and hypertext, which are far more complex and sophisticated. Furthermore, in the soft-link hypertext model, both LTM and STM are applicable, i.e., two sets of information are available for the information user.

10.2.2 My Contributions

With the soft-link hypertext model, my contributions to information retrieval area can be listed as follows:

1. Extending the idea and notation of hypertext. The definition of node, anchor, link and path are broadened and made more flexible. Some new terms, including hard link, soft link, hard path and soft path, are coined;

2. Introducing a new semantic net into the conventional global index and applying the consequential conceptual index in IR. It leads to a richer set of expressions to represent not only concepts, but the relationships between concepts;

3. Defining and developing an advanced IR model, called the soft-link hypertext model. Based on the conceptual index, the model covers the Boolean search, the conventional hard-link hypertext and soft-link hypertext into one infrastructure, and therefore, presents a superior and more flexible design for IR;

4. Designing an automatic formulation mechanism in the model for the generation of the conceptual index;

5. Pioneering in applying statistical information to improve the dynamic hypertext structure. A new set of competitive learning rules is defined;

6. Adopting inference and reasoning in hypertext to aid on-line navigation. Probability theory is used for the model to make predictions of the user's information needs;

7. Implementing such a soft-link hypertext model and running some information retrieval experiments in small-scale controlled applicational environments, which reveals some new and very interesting features about IR.

Moreover, my contribution also includes the introduction of a new set of competitive learning rules, which is similar to classical Kohonen learning laws. The two differ in that, in our system, a value attached to each weight is used to reflect the density
function of the training data; whereas in the traditional Kohonen learning, the relative positions of the weights are used to reflect this density function. Although I was not able to follow this line of research as it is not the main stream of my Ph.D work, I believe that such learning rules constitute a very interesting research topic and should have wider applications.

10.2.3 Relations with Other Recent Research Activities

Information Retrieval is a fast changing and developing area; I have, therefore, constantly compared and contrasted my work with other recent research activities.

The research presented in this dissertation started with 12 months' survey and investigation in IR. During that time, I collected and read quite exhaustively the literature on all relevant subjects. This included some 250 papers, 200 books, 60 research notes and reports, dating from 1945 to more recent ones in 1990s; they covered information retrieval, database technology, artificial intelligence, programming languages and tools. This thorough reading laid a very good foundation for my research in the following stages.

I spent the next 18 months on the development and implementation of the soft-link hypertext model. The literature published during that period describes the trend and most recent developments in the IR community. The best examples include Frisse's report on hypertext models [69], Croft's analysis of AI in information retrieval [45], Savoy's presentation of his spreading activation in IR [139], and Murray's paper on limitation and difficulties of hypertext [116]. Whilst I am glad to see that my research is addressing the issues proposed in the literature and is in accord with the recent trend of development of large-scale intelligent IR models and systems, those people's work helped me in the further refinement of the soft-link hypertext model.

By the time I had finished the first draft of this dissertation, I read the first serious report of the Dexter hypertext model (Communications of the ACM, February 1994). This led to a deeper analysis of the soft-link hypertext model and its relationship with other hypertext systems (Section 8.2). Meanwhile, I revisited all 250 papers again,
Finally, during the last 36 months, I have followed very closely with the a number of research projects on the related fields. These projects include:

1. The PODA Project, Department of Computer Science, UCL;
2. The research activities in University of Bellcore;
3. X.500 Quipu Project, Department of Computer Science, UCL;
4. The World-Wide Web (WWW) project;
5. The Wide-Area Information System (WAIS) project.

10.2.4 Limitations

Inevitably, there are a number of distinct limitations to the work that is described here, because of the restriction of time, man-power, and research scale of Ph. D.

At the broadest level, all the discussions on the soft-link hypertext model are concentrated on the Index Space, whilst no specifications are given to the Document Space and the front end (Figure 8.4). Fundamentally, a proper database model should be adopted for the information component in the model, and a more collaborative relationships among the three layers should be investigated in order to build up an ideal intelligent information model. However, I did not have enough time to look into the details on this subject.

Self-adjustment and learning have the potential for, and should be applicable to, all aspects of features in the model. Nevertheless, in this dissertation, it has only been used in improving the conceptual index and helping the information search and navigation. Further possibilities are discussed in Section 10.3.1.

During the limited specific research period of my Ph. D, the main concern has been to construct the model, highlight its dynamic and self-adjustable features, and have a small-scale implementation and demonstration of the methodology so that its potential can be shown. Therefore, all other work, including the actual implementation of the Enhanced SuperBook, construction of stoplist, employment of morphology during conceptual indexing and formulation, etc. is reduced to the minimum. Whilst the positive aspect of this is that I could reach some tangible research results during the 36 month Ph. D work, the negative side is that these questions, including the aspects mentioned above, remain unanswered.
10.3 Future Work

10.3.1 Further Development of the Model.

Although the soft-link hypertext model proposed in this dissertation has shown its potential, it can be improved further in at least the following four aspects:

1. The strength of the soft-link hypertext model could be further enhanced with a good database model adopted on its Document Space and an effective specification defined on its front end, so that the three layers could collaborate more harmonically;

2. As each of the conceptual relation (e.g. model, definition, etc.) is a concept itself, this dual role could make it possible to access a destination anchor (composite concept) by two approaches, e.g., definition of polymer could be approached via polymer → definition of polymer, or definition → definition of polymer. This could augment the model for both index formulation and navigation;

3. The learning mechanism used in the model can be applied more extensively, in self-adjustment of stoplist, self-adjustment of formulation rules, and automatic assessments of collected browsing trails;

4. The Boolean search could be applied even more extensively in the soft-link hypertext model. For example, for a hypertext source anchor (e.g., polymer), the global search would be useful in helping to search for a requested soft link.

10.3.2 Large-scale Prototype and Evaluation.

If a large-scale soft-link hypertext prototype can be set up publicly and operated continuously for a period of time (e.g., 6 - 12 months), a full range of evaluation of the model could be made possible. This should reveal some very interesting results. After some lengthy exposure to public access by a large number of information users who approach the system in a natural environment with actual information needs, a group of useful soft links could become prominent. This group could either turn into permanent links (hard links), or provide a pool from which the hard links can be set up by domain experts. Then, conducting a large-scale user survey (like Experiment 4 proposed in this dissertation) to investigate the applications of different IR models
should reveal some positive information for the soft-link hypertext model. Such kind of pilot should be really necessary and helpful for the extensive application and commercialization of the model.

10.3.3 Improvement of the Existing IR System.

It should not be very difficult to apply the soft-link hypertext model to a wide range of existing hypertext systems. I have shown that the model by itself shares the fundamental structure with the standard Dexter model. This means that the soft-link hypertext model is, in fact, in accord with the present hypertext systems such as NoteCards, Neptune, KMS and Augment. Particularly, it addresses the same issues as current reconfigurable hypertext systems including Microcosm, Hyperform, and Intermedia. Should a semantic net be available and serve with the original global index structure to facilitate a conceptual index, the soft-link hypertext model could be applied immediately to these systems, either as the mainstream, or in cooperation with the existing structure. Such a practice could dramatically increase the user's mobility in a large-scale and distributed Information Space, enhance the system's various features and improve the IR services. The actual implementation in this dissertation (the Enhanced SuperBook which is built on the SuperBook) has already demonstrated the possibility of such an extension, as well as the benefits thereafter. Moreover, the consequential system still constitutes an open configuration, which can be extended even further.

A good example could be the applications of the soft-link hypertext model on Microcosm, which is a hypermedia system developed in University of Southampton. Whilst pioneering in the exploration of open and reconfigurable hypertext, Microcosm still supports the standard Dexter hypertext model. Its exceptional features include a configurable Index Space (which is a database of links called “linkbase” [88]), a series of filters for the manipulation of the Document Space and the linkbase, and elegant specifications of the operations and applications of the filters. The essence of Microcosm is to provide the system users with the ability to tailor the functionality of the system to meet their own requirements and extend further the present hypertext/hypermedia models and systems to extensive electronic information services. The soft-link hypertext model developed in this dissertation can be well applied to Microcosm. The adoption of the conceptual index, as well as its automation generation and self-adjustable attributes, in the Index Space (linkbase) could enhance
further the manipulatable and reconfigurable features of Microcosm. Meanwhile, the splendid specifications defined between layers and the operations of filters in Microcosm could complement the defective side of the soft-link hypertext model and constitute a good example for the further development of the model (Section 10.3.1). Such an extension on both models could be very useful and stimulating for the recent hypertext community.

10.3.4 Impact on Future Generations of IR

The soft-link hypertext model, as well as the ideas behind it, should have a great impact on the further development of information system. With the increasing demand of large-scale distributed information system and decreasing cost of machine power (processing speed and memory), more and more applications of automatic mechanisms, machine intelligence, self-evolution and self-adjustment methodology in IR are inevitable. In this dissertation, I have demonstrated its potential and initial benefits via a small-scale research model and prototype, and have shown the possibilities of further development. It is confidently expected that such a methodology will be applied more and more extensively in the coming generations of global information systems.

Furthermore, although this research is text-based, its methodology, the state-space and operations defined in the soft-link hypertext model, the activation spreading and learning rules, and experiences obtained via implementing and assessing the intelligent systems are also applicable for hypermedia and multimedia systems.

10.4 Conclusion

In this dissertation, a new information model, called the soft-link hypertext model, is proposed and developed. The main features of the model include automatic generation of information structure, self-adjustment and evolution based on the statistical information, applications of machine intelligence in accumulating the past experiences, inferring a better user model and reasoning, multiple accessments to the information stored, effective on-line navigation, and user-centered services for information search and retrieval.
Still maintaining the traditional Dexter three layer model on a high level, the soft-link hypertext model extends the idea of hypertext into a new dimension and facilitates a much more powerful conceptual index structure. The conceptual index structure integrates the conventional global index with a special semantic net. As a result, a much richer set of concepts, as well as the relationships between concepts, can be represented in the data structure. Based on such a data structure, the soft-link hypertext model is able to cover several IR methodologies, including the Boolean search, the probability model, the traditional hard-link hypertext model and the soft-link hypertext, in one infrastructure.

Applications of the soft-link hypertext model in IR leads immediately to the following three main benefits. First of all, it becomes possible to formulate automatically the conceptual index structure needed for more sophisticated IR environment such as hypertext. Secondly, various kinds of intelligent methodologies can be applied in the data structure for self-adjustment and evolution. Finally, more auxiliary information becomes available for assistance during information retrieval. It is obvious that the issues addressed by the soft-link hypertext model are exactly the bottlenecks for the present IR and represent the main elements for future generations of IR systems.

The state-space and all operations happening on the model are defined in this dissertation. It is implemented in a soft-link hypertext system and extensively assessed at different levels. It has been shown that such a system could increase the user’s mobility in the Information Space and increase the effectiveness and efficiency of information retrieval.

Finally, such a soft-link hypertext model opens many new lines for further developments in IR. Its enormous potential should have a great impact on future generations of global intelligent information systems.
# Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>bibliography</td>
<td>Bibliographies are pointers or links to documents. [127]</td>
</tr>
<tr>
<td>browsing</td>
<td>The art of not knowing what you want until you find it. [127].</td>
</tr>
<tr>
<td>cognitive overhead</td>
<td>The additional effort and concentration necessary to maintain several tasks or trails at one time.</td>
</tr>
<tr>
<td>document</td>
<td>A structured body of natural language and graphics that carries a cohesive message. [125]</td>
</tr>
<tr>
<td>Document Space</td>
<td>Including all information components (documents) in an information system. The Document Space constitutes the main source of useful information for the information user.</td>
</tr>
<tr>
<td>index</td>
<td>Another representation of the Document Space [127]. It provides a mapping from an index term to the related documents where the term is included.</td>
</tr>
<tr>
<td>indexing</td>
<td>Generation process for an index</td>
</tr>
<tr>
<td>index language</td>
<td>Any instrument for the organization, description, and retrieval of knowledge that consists of verbal or notational expressions for concepts and their relationships and which displays these elements in an ordered way. [127]</td>
</tr>
<tr>
<td>Index Space</td>
<td>An auxiliary structure outside the Document Space in an information system. It usually includes an index and some necessary tools needed for storing, organizing and accessing the information components in the system. Its operation is to make information search and retrieval more efficient.</td>
</tr>
<tr>
<td>information retrieval</td>
<td>Dealing with the representation, storage, organization, and accessing of information items. [134]</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Information Space</td>
<td>Information Space denotes the union of the Index Space with the Document Space. It represents the whole database of an information system.</td>
</tr>
<tr>
<td>intelligent IR system</td>
<td>a computer system with inferential capabilities such that it can use prior knowledge to establish, by plausible reasoning, a connection between a user’s probably ill-specified request and a candidate set of relevant documents. [20]</td>
</tr>
<tr>
<td>model</td>
<td>a representation, usually on a smaller scale, of a device, system, structure, etc.</td>
</tr>
<tr>
<td>precision</td>
<td>the ratio of the relevant information items retrieved by an information system to the total number of items retrieved.</td>
</tr>
<tr>
<td>recall</td>
<td>the ratio of relevant information items retrieved by the system to the total number of relevant items in an Information Space.</td>
</tr>
<tr>
<td>searching</td>
<td>comparing and matching the parsed form of a query with those of documents. [127]</td>
</tr>
<tr>
<td>search formulation</td>
<td>moving a query into another/simpler representation which is understandable by an information system. [127]</td>
</tr>
<tr>
<td>self-organizing</td>
<td>self-organization process (or topology-preserving mapping) is to transform a signal pattern of arbitrary dimensionality onto a one- or two-dimensional array. [96]</td>
</tr>
<tr>
<td>semantic net</td>
<td>the abstraction of a document as a network of concepts and relations is a semantic net. [127]</td>
</tr>
<tr>
<td>term</td>
<td>a word or sequence of words that refers to a concept. [127]</td>
</tr>
<tr>
<td>text</td>
<td>any recorded body of information. [125]</td>
</tr>
<tr>
<td>thesaurus</td>
<td>a set of concepts in which each concept is represented with at least synonymous terms, broader concepts, narrower concepts, and related concepts. [127]</td>
</tr>
</tbody>
</table>
REFERENCES


Appendix A

Summary of Schemas Used in the Dissertation

1. Schema IndexSpace defines the state-space of the soft-link hypertext model;

\[
\text{IndexSpace} \\
\begin{align*}
\text{Concept:} & \mathbb{P} \ \text{CONCEPT} \\
\text{ConceptualRelation:} & \mathbb{P} \ (\text{CONCEPT} \times \text{CONCEPT}) \\
\text{ActivationLevel:} & \text{Concept} \rightarrow A_C \\
\text{Weight:} & \text{ConceptualRelation} \rightarrow W_R \\
\text{ConceptVisisted:} & \text{Concept} \rightarrow N_C \\
\text{ConceptualRelationVisited:} & \text{ConceptualRelation} \rightarrow N_R \\
A_C : & (0, 1) \\
W_R : & (0, 1) \\
N_C : & \mathbb{F} \ \mathbb{N} \\
N_R : & \mathbb{F} \ \mathbb{N}
\end{align*}
\]

\[
\text{dom ActivationLevel} = \text{dom ConceptVisisted} = \# \text{Concept} \\
\text{dom Weight} = \text{dom ConceptualRelationVisisted} = \# \text{ConceptualRelation} \\
\forall w_{ij} \in W_R: w_{ij} = \frac{1 + n_{ij}}{d_C(i) + n_i}
\]

2. Schema InitIndexSpace describes the operation before index formulation starts;

\[
\text{InitIndexSpace} \\
\Delta \text{IndexSpace} \\
\text{Concept} = \phi
\]
3. Schema \textit{AddConcept} describes the operation of global indexing;

\begin{verbatim}
AddConcept
\[ \Delta \text{IndexSpace} \]
\[ V_{Ci}?: \text{CONCEPT} \]
\[ V_{Ci} \not\in V_C \]
\[ \text{Concept}' = \text{Concept} \cup \{V_{Ci}\} \]
\[ d^+_C(i) = 0 \]
\[ \text{ActivationLevel}' = \text{ActivationLevel} \cup \{V_{Ci} \mapsto 0\} \]
\[ \text{ConceptVisited}' = \text{ConceptVisited} \cup \{V_{Ci} \mapsto 0\} \]
\end{verbatim}

4. Schema \textit{AddConceptualRelation} describes the operation of hyperization;

\begin{verbatim}
AddConceptualRelation
\[ \Delta \text{IndexSpace} \]
\[ V_{Rij}?: (\text{CONCEPT} \times \text{CONCEPT}) \]
\[ V_{Ci} \in V_C \land V_{Cj} \in V_C \land V_{Rij} \not\in V_R \]
\[ \text{ConceptualRelation}' = \text{ConceptualRelation} \cup \{V_{Rij}\} \]
\[ \text{ConceptualRelationVisited}' = \text{ConceptualRelationVisited} \cup \{V_{Rij} \mapsto 0\} \]
\[ d^+_C(i) = d^+_C(i) + 1 \]
\[ \forall w_{ik} \in W_R : w_{ik}' = \frac{1+n_{ik}}{d^+_C(i)+n_i}, \quad (k = 0, 1, ..., d^+_C(i) \text{ and } k \neq j) \]
\[ \text{Weight}' = \text{Weight} \cup \{V_{Rij} \mapsto \frac{1}{d^+_C(i)+n_i}\} \]
\end{verbatim}

5. Schema \textit{RemoveConceptualRelation} describes the operation of deleting soft links;

\begin{verbatim}
RemoveConceptualRelation
\[ \Delta \text{IndexSpace} \]
\[ V_{Rij}?: \text{ConceptualRelation} \]
\[ \text{ConceptualRelation}' = \text{ConceptualRelation} \setminus \{V_{Rij}\} \]
\[ \text{ConceptualRelationVisited}' = \text{ConceptualRelationVisited} \setminus \{V_{Rij} \mapsto n_{ij}\} \]
\[ \text{Weight}' = \text{Weight} \setminus \{V_{Rij} \mapsto w_{ij}\} \]
\[ d^+_C(i) = d^+_C(i) - 1 \]
\[ n_i = n_i - n_{ij} \]
\[ \forall w_{ik} \in W_R : w_{ik}' = \frac{1+n_{ik}}{d^+_C(i)+n_i}, \quad (k = 0, 1, ..., d^+_C(i)' \text{ and } k \neq j) \]
\end{verbatim}
6. Schema *RemoveConcept* describes the operation of deleting index terms;

\[\begin{align*}
\text{RemoveConcept} \\
\Delta \text{IndexSpace} \\
V_{Ci} &: \text{CONCEPT} \\
\end{align*}\]

\[\begin{align*}
V_{Ci} \in V_C & \land \forall V_{Cj} \in V_C : (V_{Rij} \in V_R \lor V_{Rji} \in V_R) \\
Concept' &= Concept - \{V_{Ci}\} \\
\text{ActivationLevel}' &= \text{ActivationLevel} - \{V_{Ci} \mapsto a_i\} \\
\text{Concept Visited}' &= \text{Concept Visited} - \{V_{Ci} \mapsto n_i\}
\end{align*}\]

7. Schema *SessionStarts* describes the operation that the model is ready for application;

\[\begin{align*}
\text{SessionStarts} \\
\Delta \text{IndexSpace} \\
PATH' &: \text{IF CONCEPT} \\
\forall a_i \in AC: a_i &= 0 \\
PATH' &= \emptyset
\end{align*}\]

8. Schema *ForwardSpreading* describes the applications of the soft-link hypertext model in information retrieval;

\[\begin{align*}
\text{ForwardSpreading} \\
\Delta \text{IndexSpace} \\
\text{PutInOrder}: (\text{Concept} \times \text{decoder}[AC]) & \to \text{seq Concept} \\
PATH' &: \text{IF CONCEPT} \\
V_{Ci} &: \text{CONCEPT} \\
\text{repl} &: \text{seq Concept} \\
\end{align*}\]

\[\begin{align*}
V_{Ci} \in V_C: a_i' &= a_i + 1 \\
\forall V_{Cn}: (V_{Cn} \in V_C \land V_{Cn} \not\in PATH') & : a_n' = a_n + \sum_m w_{mn} \times a_m \\
\forall V_{Cn} \in PATH': a_n' &= 1 \\
PATH''' &= PATH' + \{V_{Ci}\} \\
\text{repl} &= \text{PutInOrder} (\text{Concept} \times \text{decoder}[AC])
\end{align*}\]
9. Schema *BackwardLearning-1* describes the learning process when a concept is visited by the user:

\[
\text{BackwardLearning-1}
\]
\[
\Delta \text{ IndexSpace}
\]
\[
V_{C_i}?: F \text{ CONCEPT}
\]
\[
V_{C_i} \in V_C \wedge a_i = 1
\]
\[
\wedge \exists V_{Rij} \in V_R:
\]
\[
n_{ij} = n_{ij} + 1
\]

10. Schema *BackwardLearning-2* describes the learning process when a conceptual relation (soft link) is visited by the user.

\[
\text{BackwardLearning-2}
\]
\[
\Delta \text{ IndexSpace}
\]
\[
V_{C_j}?: F \text{ CONCEPT}
\]
\[
V_{C_j} \in V_C \wedge a_i = 1 \wedge a_j = 1
\]
\[
\wedge \exists V_{Rij} \in V_R:
\]
\[
n_{ij} = n_{ij} + 1
\]
\[
\forall w_{ij} \in W_R: w_{ij}' = \frac{1 + n_{ij}}{d_{ij}^2(i) + n_i}
\]
Appendix B
The List of Information to Be Retrieved in Experiment 2

1. cleaning agent
2. department of biology
3. metal bonding
4. carbon dioxide
5. physical chemistry
6. water condensation
7. diamond cutting (x)
8. chlorine dioxide
9. fiber-optic materials
10. 1.8mm film
11. collision gas
12. immune suppression
13. chronic inflammation
14. metabolic inhibitors
15. water molecule
16. aging phenomenon
17. synthetic polymer
18. temperature sensor
19. excessive temperature
20. thermal conversion
21. atomic absorption
22. amorphous phase
23. cohesive strength (x)
24. density fluctuation (x)
25. boltzman distribution
26. cartridge elusion
27. frequency domain
28. silicate hydration
29. joint variability
30. methyl nitrite
31. nucleation rate (x)
32. physical chemistry
33. radius of gyration (x)
34. aromatic resin
35. rubber compound (x)
36. scattering cell (x)
37. temperature gradients
38. tracer solutions
39. recovery of uranium
40. vinyl bromide

(x) means the information is not included in the Information Space.
Appendix C
Questions for Experiment 3

Air pollution damage to museum collections has emerged as a major issue in art conservation. Outdoor air pollution, and more specifically, the widespread damage caused by acid deposition on stone, has been recognized as a threat to cultural property for a number of years. Furthermore, damage to works of art due to exposure to indoor air pollutants has also received increasing attentions recently.

Accordingly, researchers have undertaken a series of surveys to investigate the selected air pollutants in several museums in the United States. Some of the research results are published in the journal, Environmental Science and Technology. They reveal some interesting facts about the indoor/outdoor air pollution and damages to museum collections...

Do read on to find more information about the surveys!
• Where are the surveys conducted?

• How many museums have been investigated?

• What is HVAC?

• Which criteria are used for the selection of museums?

• Which model of electron capture detector is mentioned?

• Which air pollutants are investigated in these surveys?

• Following the introduction of air pollution controls in the 1960s, Did the level of sulfur dioxide drop? To how much?

• In the United States alone, how many pounds of toxic chemicals were discharged into the air in 1987, according to the report?

• Where is the chlorinated hydrocarbons from?

• What do you know about the retention time?
1. Which substance(s) are there on the outer surface of the hamburger, which would be vaporized during its cooking process?

(a) Fine organic carbon particles;
(b) Fatty acids;
(c) The normal alkanes with carbon \( C_{21} \);
(d) I do not know.

2. In Hisham’s survey about the air pollution in southern California museums, which of following organization(s) have been investigated?

(a) An Archeological museum;
(b) An ethnology museum;
(c) A natural history museum;
(d) Several major art museums with many types of collections;
(e) One of the largest libraries in the United States;
(f) Several historical buildings;
(g) All of the above;
(h) I do not know.
3. For more than 25 years it has been recognized that rice fields are a large source of atmospheric methane. Compared with those in China, how are the methane emission rates from rice fields in the United States and Europe?

(a) About the same level;
(b) 23 times higher;
(c) 4 - 10 times lower;
(d) I do not know.

4. Portable kerosene heaters are often used as space heaters in homes to reduce energy costs in the United States. However, when unvented kerosene heaters are used in homes with small air volumes and low air exchange rates, such as mobile homes, high concentrations of air pollutants can result. Researches have been conducted to assess human exposure to air pollutants from unvented kerosene heaters in mobile homes. Which pollutants have been identified in those experiments?

(a) Presence of unburned kerosene fuel;
(b) Increase of CO and organic levels;
(c) Airborne particles of < 10 um in diameter;
(d) I do not know