# INFORMATION RETRIEVAL AND HYPERTEXT

EDITED BY

**Maristella Agosti**
*University of Padua*
*Italy*

■

**Alan F. Smeaton**
*Dublin City University*
*Ireland*

KLUWER ACADEMIC PUBLISHERS
Boston/London/Dordrecht

**Distributors for North America:**
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA

**Distributors for all other countries:**
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS

**Library of Congress Cataloging-in-Publication Data**

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

*Printed on acid-free paper.*

Printed in the United States of America

# CONTENTS

PREFACE ix

Part I BASICS 1

1 AN OVERVIEW OF INFORMATION RETRIEVAL
*Alan F. Smeaton* 3
1 Information Retrieval and Information Retrieval Tasks 3
2 Indexing and Retrieval 5
3 Information Retrieval Implementation 8
4 Evaluation in Information Retrieval 9
5 Mathematical Modelling and Logic in Information Retrieval 11
6 Natural Language Processing in Information Retrieval 13
7 User Issues in Information Retrieval 15
8 Miscellaneous Issues 17
9 Conclusions 20

2 AN OVERVIEW OF HYPERTEXT
*Maristella Agosti* 27
1 Introduction 27
2 Hypertext Basic Concepts 28
3 Relevant Features of Hypertext Systems 29
4 Towards Hypertext and Information Retrieval 32
5 Hypertext Information Retrieval 34
6 Design and Development of HIR Systems 35
7 Networked HIR 42

# 2

# AN OVERVIEW OF HYPERTEXT

**Maristella Agosti**

*University of Padua (Italy)*

## 1 INTRODUCTION

In the late eighties, the first generation of hypertext systems started to become widely available [17, 29, 36, 38]. Right from the beginning of their availability, these systems have been a topic of great attention from the information retrieval research community as potential tools to be used to implement new information retrieval ideas and capabilities.

Some of the application areas where hypertext systems, both prototypes and experiments, have flourished include dictionaries, encyclopaedias, medical textbooks, product catalogues, help systems, technical documentation, and museum exhibits [36]. Some of the application prototypes have incorporated information retrieval ideas and capabilities; an example of such an application prototype has been designed and developed as early as 1988 by *Frisse* for medical informatics [22, 23].

The availability of hypertext systems for personal computers has made it easy to provide software tools for developing prototypes for different types of applications and in different areas. More importantly for the information retrieval area was that it was now possible to use a hypertext system to conduct experiments that previously were very difficult and time consuming because of the necessity of developing ad hoc software. Examples of such hypertext systems are *GUIDE*, made available from 1985 for Apple Macintosh and later on for the IBM PC and compatible personal computers, and *HyperCard* made available for the Apple Macintosh starting from 1987.

In this chapter we address the characteristics of hypertext systems that have made them important as general tools for managing information and, in particular, as useful tools for the management of information in information retrieval applications. We begin with an introduction to basic concepts and useful features of hypertext but the main focus of the chapter is to cover hypertext in the context of information retrieval applications including the topics of searching, dynamic or active hypertext, text to hypertext conversion, user modelling and evaluation. We also briefly look at networked systems including the WWW.

## 2 HYPERTEXT BASIC CONCEPTS

A concept that is basic to the design of a *hypertext system* is to provide the user with a tool that opens the possibility of managing a textual document in a *non-linear* or *non-sequential* way. By that we mean a software tool that gives the user the possibility of using a textual document not only in a sequential way and thus providing the capability for creating, managing and linking parts of text, to give the user a multi-dimensional document that can be used and explored by following different paths through it.

In the hypertext approach to information management, a document is separated into parts or fragments; all fragments are stored and managed in a *network* of nodes, where each node of the network contains a fragment and related nodes are connected through connections called information links. Thus the network of connections gives the possibility of following a path through nodes and seeing fragments that are related together. Each sequence of connections forms a different path to fragments of the overall document.

The *hypertext* is the combination of the fragments of the original flat document together with the connections between those fragments, where each fragment is managed in a *node* and each connection between fragments through a *link*. Thus:

**hypertext = nodes + network of links.**

Since hypertext systems were initially developed as supporting tools for different kinds of documentation, the original idea of a hypertext system as able to manage only one single document has been extended to that of a system capable of managing a *collection of documents*. A hypertext system allows authors or groups of authors to link documents together, create paths through a collection of related documents, and create references that point from one document to external documents which are associated with it [40]. Thus the document collection managed by a hypertext system can be just a single large document (like a complete book), a number of large documents, or many medium/small sized documents related together for some reason, and thus it makes sense to combine and present them to the user as a single unique *hypertext document*. What this means is that in the hypertext approach to information management, the collection is stored and managed as a *network* where each node of the network contains a part of a large or medium-sized document, or contains a complete small document. From the original idea of storing and managing only textual documents, the capabilities of hypertext systems have been extended to those of systems able to handle any kind of media that can be digitised on a computer. Because of this improvement in capabilities, the term *hypermedia* is now in common use for a hypertext where the managed fragments of documents are digitised forms of different media and not just of text. Since most present-day hypertext systems are able to manage different kinds (media) of digitised documents, the term *hypertext* is used in this chapter indifferently for *hypertext* and *hypermedia*.

Examples of collections of documents that could be managed by a hypertext system would include an entire encyclopaedia, the digitised version of the paintings of a museum together with textual descriptions of them and maps to locate them in the museum, the medical records for a group of patients, or a collection of bibliographic documents together with reviewer comments on them and citation lists.

## 3 RELEVANT FEATURES OF HYPERTEXT SYSTEMS

The passage from a flat collection of documents to a hypertext requires the design of the target hypertext application. The initial document collection needs to be fragmented and the links between fragments need to be built. The process of passing from a flat collection of documents to a hypertext is referred to as *authoring* the hypertext. To perform the authoring of a hypertext it is necessary to use a hypertext application design method, and a set of tools for editing the fragments, similar to an interactive browser which permits the interactive examination of the nodes, and a tool similar to an editor for preparing and editing the different types of multimedia nodes [14, 28].

The hypertext system is the tool which can be used to manage the fragments of the document collection and the links between them. The nodes and links constitute the hypertext which can also be thought of as a hyperdocument. The difference between the hypertext system and the hypertext is the same difference as there is in the database management area between the database management system, i.e. the software tool that manages the database, and the database, i.e. the managed information base.

Some hypertext concepts can now be examined in more detail:

- **Node**: A node contains any kind of digitised data that can be managed and presented through a computer using its screen and any other output device. It is a fragment of text, a graph, a drawing, a piece of music or in general any kind of audio, a video sequence, an animation clip, other possible type of data, or indeed any useful combination of these. Some *aggregation mechanisms* may be available in the hypertext system to manage some aggregations of nodes; for example, it may be possible to have a *composite node* as a node that manages a collection of related nodes.

- **Link**: A link implements a logical connection between two related nodes; the *origin* is the node from which the logical connection between two fragments emanate; the *destination* is where a connection between nodes ends. Two nodes which have to be explored or viewed sequentially are connected by a link. Each sequence of nodes connected by links constitutes a possible exploration path of the hypertext. Different types of links have been defined depending on the functionalities that need to be implemented by a hypertext. The most relevant of these to information retrieval applications are addressed in the following chapters of this book. Here it is important to recall the difference between explicit and implicit links; an *explicit link* is a link that makes available an explicit reference between two nodes; explicit links are built during the authoring process and they constitute the main part of the hypertext network. An *implicit link* is a link that is implicitly present and starts from a node; an implicit link can be activated using a word present in a node: if a user asks to see all nodes that contain a word and all nodes containing the word can be made available to the user, then these implicit links are usually activated at run time and correspond to links not specifically created during the hypertext authoring process.

- **Anchor**: A node may have several out-going links and in such cases each link is associated with a small part of the node; this small part is named an *anchor* of the node. When the user activates an anchor and follows the associated link, he navigates the hypertext network.

The nodes and links of a hypertext can be created or deleted; the information contained in a node and the links between nodes can be modified so the structure and the contents of a hypertext can evolve dynamically.

The network of links constitutes the only structure which can be used to *navigate* in the hypertext. To navigate the hypertext, the user needs a tool able to follow the links and such a tool is called a *browser*. A browser usually incorporates both navigation and browsing facilities. It is important to note that if a link does not exist between two nodes which are semantically related, they cannot be viewed (retrieved) by a user who is browsing the hypertext. The only way in which two or more related nodes that have not been explicitly connected by links can be retrieved is by searching the network for some word, string, keyword or attribute value which nodes have to share. In effect this makes use of an implicit link between nodes. Normally it is only one specific and exact string, keyword or attribute value that can be used for searching in such circumstances. At the present time of the evolution of hypertext systems this feature cannot be related to the exact match retrieval techniques [10] which use a query language based on Boolean logic, and are available in most of the present operational information retrieval systems.

The *user interface* is a crucial part of a hypertext system because it is through this that the user can or cannot reach useful information. Two of the major problems the user interface of a hypertext system needs to address are the *disorientation* and the *cognitive load* of the user. These can be defined as:

- **Disorientation** is a user's feeling when he or she does not fully understand what are the navigation and browsing facilities made available by the hypertext system; in this case the user is "lost" in the hypertext and does not know in what way to make use of the navigation and browsing features. Most hypertext systems have browsing aids to assist the user in navigation; all systems have a backtrack capability to assist the user in the backward reconstruction of the exploration path constructed using the hypertext.

- **Cognitive load** is the effort the user has to face to understand and learn the "cognitive model" that has been used in designing the hypertext application and the hypertext system itself. The information base managed by the hypertext system will have been structured during the authoring process with a specific cognitive model in mind and if this model is not made clear to the user, the cognitive load of the user can be too large and the user can become overloaded. The ability of the user to use the hypertext is directly related to his knowledge of the hypertext structure being used.

The simplicity and consistency of the hypertext structure help in reducing the load on the user and also in reducing the learning time necessary to reach a sufficient level of knowledge of the hypertext.

## 4 TOWARDS HYPERTEXT AND INFORMATION RETRIEVAL

The models of information retrieval which have been the foundation of present information retrieval systems are based on the assumption that the documents to be managed are linear. Operational information retrieval systems are based on retrieval techniques which answer a user's query with a set of documents, so all documents are on the same "plane". In hypertext, the documents are not on the same plane. When a user uses a hypertext, he personally visualises the links between documents and he can see the number of nodes which constitute the path or distance between two documents. Furthermore, it is always possible in a hypertext to implement the direct connections between a document and all the other documents which are referenced by it. This possibility is not at all obvious in operational information retrieval systems. The network of links which connect documents and fragments of documents in a hypertext can perform the same task the *indexing term structure* is performing in an information retrieval system.

The problem of the representation of the semantic content of documents is central to the information retrieval area, but has only initially been addressed in the design of the first generation of hypertext applications where it was more urgent to address and solve issues related to the structure, management, and presentation of documents. The network of links of the hypertext has been semantically enriched by the designers of the second generation of hypertext applications, because this is the clue for the development of hypertext systems able to present and make use of the *content* of the managed information. It is the network of links that can semantically guide the user in browsing the hypertext and hyperdocument, thus the network of links has to support some sort of vocabulary control [25]. This is because one objective of vocabulary control is to make a comprehensive search on a specific topic easier by linking together terms whose meanings are related.

The aspect of hypertext technology that is so important for the retrieval of information is the ease of linking different pieces of information, possibly presented in different media. In fact the media used in fragments of information that must be handled in hypertext information retrieval applications are very different from each other. Some examples of these are:

- fragments of textual documents e.g. the abstract of a bibliographic reference;
- structured data similar to the data managed by database management systems applications e.g. date of publication of a document;
- a list of terms that represents the information content of a document e.g. a list of terms from a thesaurus used in the description of the information content of documents of a collection;
- the definition of a term used by the indexing term's structure as used when representing the semantic contents of the documents of a collection e.g. the definition of a term in a thesaurus that is integrated in the application.

Thus, using the capability of easy linking between different pieces of information, it is possible to:

- link a document with a term that represents some aspects of the information content of the document itself;
- connect two related documents for example, for bibliographic collections, this feature gives the possibility of connecting one bibliographic reference to one of its cited references, and vice-versa;
- relate a term to a node containing its definition and use;
- link two related terms.

The capability of easy linking of different pieces of information, which is considered very important in the development of effective hypertext information retrieval applications, can produce information retrieval hypertexts that are very difficult to use for the end-user because these same capabilities can generate user disorientation and cognitive overload. To make use of the specific capability of hypertext systems together with information retrieval operations, work has started in the new area of *hypertext information retrieval.* Some of the initial papers addressing the issues related to the combination of hypertext and information retrieval capabilities to produce a new kind of innovative information management tools are [2, 3, 18].

## 5 HYPERTEXT INFORMATION RETRIEVAL

The collection of documents which is stored and managed by an information retrieval system is usually large and can range from thousands to millions of documents. The collection can be imagined as a set of documents on a flat surface; each document in the collection is composed of linear parts of text. As a result of this organisation, when the user makes an inquiry against the collection of documents, all documents are equally potentially relevant to a query.

The indexing term structure is the data which is associated with each document for representing its semantic content. The indexing term structure is used to select and retrieve documents during query processing. Because of this, indexing terms constitute access points to the collection. One specific information retrieval model may differ from another in the kind and structure of indexing terms which are used during system operations.

Two parts of a hypertext can be used for the same query, even if they are implemented in a completely different way; the collection of documents managed through nodes and the network of links which connect the documents of the collection. This network is the structure that can be used to connect semantically or structurally related documents or fragments.

Activities in the field of *hypertext information retrieval (HIR)* focus basically on the design and implementation of systems capable of providing the end-user with the properties of the visionary Memex device and system of *Bush* [16]. These properties imply the design and implementation of a system capable of providing the end-user with at least these capabilities:

1. storage of a large collection of textual and multimedia documents;
2. build up of a network of semantic relationships among the multimedia components of this database.

The storage and automatic construction of such a database directly from the components of the documents of a network of semantic associations between pieces of information will give the hypothetical end-user access to a large depository of knowledge for reading, browsing and retrieving. The end-user of such a system would be given the possibility to satisfy his information needs by using, concurrently, different *retrieval techniques* based upon:

- **value**: a technique present in a highly specialised way in the majority of existing *database management* and *information retrieval systems*;
- **content**: a feature chiefly available in operational *information retrieval systems*;
- **direct association**: a possibility for the direct presentation through connecting capabilities of information on different forms of media as in available *hypertext systems*.

## 6 DESIGN AND DEVELOPMENT OF HIR SYSTEMS

Different aspects need to be addressed when the functionalities and capabilities of both hypertext and information retrieval systems are to be combined together to offer to the end-user the possibility of navigating, browsing, and searching a large collection of documents to satisfy an information need. To satisfy such information needs a HIR system can be made available to the user as a tool that combines both the searching facilities of an IR system together with the navigation and browsing facilities of a hypertext system.

Aspects to be considered in the design and construction of efficient hypertext information retrieval systems are:

- navigation and browsing *versus* direct search; this requires new retrieval models;
- the possibility of modifying the status of a hypertext from passive to active;
- automatic authoring and construction of the hypertext;
- user modelling and interfaces;
- new methods and techniques for the evaluation of HIR systems.

In the following sub-sections of this chapter, these aspects are addressed individually.

## 6.1 Navigation and Browsing *versus* Direct Search

Information retrieval modalities provided by hypertext systems are different from those of traditional information retrieval systems in that information searching is conducted by navigation through the information base and not by direct search by means of a search language. One argument for preferring navigation to searching is the possibility the user is given to dynamically construct an information path by browsing through pieces of the information base. An argument against this form of information seeking is that this method can be very time-consuming and poorly organised if the information base consists of a large collection of multimedia fragments of documents.

Most of the work in HIR has been devoted to the presentation of new retrieval models able to combine both navigation and direct search features. Some of these new retrieval models like the one presented by Chiaramella and Kheirbek later on in this book, incorporate the facility of managing the representation of objects at different levels of abstraction, whereby "object" is meant any elements of the HIR; examples of managed objects would be nodes and links. This facility provides for the creation of nodes and links which form structures allowing different levels of search depth. An example of such a structure is the *composite node* which is derived from the application of the aggregation mechanism to a collection of related nodes and gives the possibility of managing and using a collection as a single node.

In the remaining part of this sub-section relevant work for giving the user a HIR system with navigation and browsing capabilities together with direct search retrieval features are presented.

In [20] and later in [21] a combined model of IR which encompasses the principles and benefits of both free text retrieval and hypermedia is presented, as a hybrid approach combining the capabilities of browsing and querying to retrieve information from a large textual and multimedia collection of documents. Contextual information is used for the retrieval of non-textual documents. This model gives the users access to large document bases with limited structure which can be browsed whatever its topology is. The model approximates the content of documents that cannot be directly retrieved by content (e.g. images), and in fact it makes use of contextual information extracted for this purpose from a hypermedia network. Moreover, this model has been only partially evaluated because of the lack of methods for HIR system evaluation, however, experiments and observation of a prototype system have shown that the use of context information from hypermedia networks to retrieve non-textual nodes by querying is effective.

An early paper by *Lucarella* on a highly connected structure of hypertext to be exploited as a knowledge base is [26]. A complete model and architecture of a prototype that combines query-based and browsing-based retrieval methods is later presented by Lucarella and Zanzi [27] and some further work is presented as a chapter in this book. This model is based on plausible reasoning, and the hypermedia collection of documents works as an inference network. Within this model, links are labelled by the name of the relationship that exists between the two connected nodes, and a weight is associated with a link to express the strength of the relationship; this is one way of modifying the status of a hypertext from a passive to an active one. Experimental results give some insights into the capabilities of the model.

*Croft* and *Turtle* in [19] uncover the relationship which exists between information contained in hypertext links and improvement of retrieval effectiveness. The results are used to develop a new HIR model which also has the capability of enabling the automatic construction of links.

*Bruza* and *van der Weide* generalise a two-level approach for hypertext information retrieval systems into stratified hypermedia structures in [15]. This is a general framework in which a number of approaches, such as state-of-the-art hypermedia, documents and keyword-based systems, can be considered. Furthermore, the stratified hypermedia architecture based on these structures constitutes an integration between logic-based information retrieval and the two-level hypertext approaches. This integration is realised by considering the retrieval process as navigation between layers. The hyperindex structure that is derived by applying this approach facilitates the process of query formulation.

In [4] an architecture and a new functional model have been introduced to overcome major limitations of hypertext systems in relation to IR operations. The model has been named EXPLICIT, because its main focus is on the *explicit* presentation to the user of the network of index terms and concepts that are used for the representation of the document collection. EXPLICIT incorporates some important IR functions and assists the end-user by means of a new type of associative IR. The two most important features of this a model are a semantic association [6] and an associative reading function. The possibility of using different techniques for semantic representation of the information being administered are exploited by EXPLICIT, and in the prototypes developed so far the opportunity to choose between different indexing techniques is given to the user and this, in turn, means the availability of different semantic interfacing capabilities for those users having different knowledge levels of the specific field covered by the managed document collection.

## 6.2 Status of the Hypertext

Hypertext information retrieval applications have the possibility of modifying the status of the hypertext from passive to active. This can be accomplished in different ways, for example, by attaching a kind of "level of importance" to a link in order to provide different link relevances depending upon the path the user is taking through the hypertext. In this way the user could be advised to follow one path instead of another, depending on the previous path taken from a node or from a certain anchor. This possibility can be implemented by establishing different types of links or using weights on links at construction time of the hypertext. As previously reported, an example of a model that supports weights associated with links to express the strength of the relationship can be found in [27].

Another way to provide active hypertext it is through *relevance feedback*, as reported in [8]. Although it is often thought that only hypertext can provide browsing capabilities, it must be noted that the ability to move between related documents can also be provided by information retrieval systems supporting relevance feedback [39]. Unlike hypertext, which generally has fixed links, relevance feedback allows the user to dynamically create links at run time by searching for documents similar to some others marked as relevant. However, browsing by means of relevance feedback is a very complex process and most existing IR systems supporting relevance feedback do not have satisfactory user interfaces for browsing, as it has been pointed out by Aalbersberg in [1]. Only by keeping the user interface and the interaction with the user at the simplest level it is possible to effectively employ relevance feedback techniques. In addition, a user might find it very useful also to be able to browse through the indexing items (index terms, concepts, thesaurus, etc.) and this cannot easily be provided by systems using relevance feedback.

## 6.3 Automatic Construction of Hypertext

The collection of documents made available through a HIR tool is usually a "flat" collection of documents. To transform a flat collection into a hypertext, it is necessary to *author* the hypertext. This means producing fragments of documents from the original complete document(s) and building up links among

them. To have a complete authoring method, it is also necessary to have an updating authoring method.

At present, it is common practice to manually author hypertexts. If the initial collection of documents is of large proportions and also consists of multimedia documents, a completely manual authoring can be impossible to achieve. It is therefore important to have automatic techniques for segmentation of documents, tools for the automatic generation of links, and procedures for the automatic authoring of the hypertext to insert, modify, and cancel part of it over time.

*Automatic authoring* has been addressed by researchers since the earliest days of hypertext. One of the earliest works in the field of the automatic transformation of text into hypertext is reported in [24]. This work illustrates the methodology and the implementation of a technique for converting a regularly and consistently structured document into a hypertext. Regularly and consistently structured documents are those having a well-identified and fixed structure such as, for example, bibliographic references or manual pages. The resulting hypertext is made of nodes corresponding to document parts connected by means of structural links. The methodology used is based on the reasonable assumption that there is a close relationship between the physical components of a document and the hypertext nodes. From an IR point of view, such structure-based hypertextual organisation should provide a better understanding of the semantic content of documents. The authors claim that their methodology is well suited for medium-grained documents that are regularly and consistently structured, such as for example, the collection of dissertation abstracts they used for their experiments. Larger, or less regular or consistent documents would require some manual intervention to create content-based links, that is to say links not explicitly inserted in the documents and that are meant to represent semantic aboutness. It is these kinds of links that are the most interesting from an IR point of view, since in HIR systems the main concerns are semantic navigation and browsing of the document collection.

The difficulty of the automatic construction of content links has been addressed by *Salton et al.* in [32] at the beginning of work carried further into other directions like passage retrieval [33], theme extraction and text summarisation [34]. The methodology for determining content links proposed by these authors is based on the evaluation of similarity between documents and/or parts of documents (sentences). Experiments were carried out by partitioning a textbook into smaller segments and using these as nodes. This work leaves open some questions regarding the resulting hypertext, for example that it should be tested to see if it is useful for IR purposes, both from a system and user's point of view and it would also be necessary to show that the technique is effective for document collections covering heterogeneous subjects. Some of these questions are addressed in the third chapter of this book by *Salton et al.* reprinted from SCIENCE.

*Rada* in [31] addresses the combination of structural links and content links. This author distinguishes first-order and second-order hypertext. The former uses only structural links based upon the document markup and determined by the document author. Structural links of this kind include links connecting outline headings, citations, cross-references, and indices. In second-order hypertext, links are not explicitly put into the text by the author but are detected using some automatic procedure. In the work reported, second-order links were set up between index terms using co-occurrence data. The use of first and second order links in the same hypertext enables both the structural schema of the source documents, that is the document author's schema, and a second alternative schema reflecting the way index terms are distributed across the documents, to be combined. Alternative outlines are different views of the same documents that users can employ to improve their understanding during browsing since alternative outlines offer different semantic points of view on the same document. As the author suggests, some more work should be carried out to test which type of hypertext, first or second order hypertext, the user appreciates better.

The approach to automatic hypertext construction proposed in [37] is based on computing node-node similarity. This approach is different from other approaches based on IR techniques because it uses the overall hypertext topology as a decision support for link setting. A measure of the hypertext topology is used to assess the degree of hypertext compactness. The lower the number of jumps from a node to another that the user has to follow to access desired information, the more compact the hypertext. This measure of hypertext compactness is employed to decide whether a similarity-based link should be added or not. The major contribution of this work is in proposing guidelines to control the automatic construction of the hypertext. What remains to be discussed is the relevance of the hypertext topology to the hypertext effectiveness. As the author highlighted, some very compact hypertext may result in the user being disoriented because of too many links. Moreover, a compact hypertext is not always desirable; in a hypertext with a large number of links the user can be helped in browsing by providing more information about links, such as for example information about the link type.

The problem of discovering link types has been addressed by *Allan* in [9]. The proposed technique provides a way of setting up links between passages of documents. The novelty of this work is that classical IR techniques are employed to determine the type of relationship incurring in a hypertext whose nodes are topics. Allan also addressed the problem of the number of links. To reduce the number of links and to make the visualization of the resulting graph easier, some link merging techniques are suggested and described. The techniques proposed by Allan for automatic link type identification are based on values calculated from merged links. For example, to identify a summary link, we can compute the amount of unlinked text that was added to a link end-point during link merging. The author points out a few directions that should be followed by further research in automatic authoring based on IR techniques. Among these, he suggests that additional work should be done with regard to heterogeneous documents, or documents that have been written with a non-regular writing style as most proposed techniques for automatic authoring are based on the assumption that documents are quite well-segmented into passages.

Since automatic hypertext and link creation techniques are crucial for future development of hypertext especially in the context of the World Wide Web [12], its importance indicates the value of more in-depth study. This is in particular for aspects related to the automatic authoring of multimedia documents that at present remains to be solved.

## 6.4 User Modelling and Interfaces

The use of hypertext applications has shown over time the difficulty the user may have in understanding the cognitive model which has been used in the preparation of a hypertext. User modelling and interface development techniques are necessary for building effective interfaces for the use of a hypertext.

*Belkin*, *Marchetti* and *Cool* present in [11] the design of an interface supporting BRowsing And QUEry formulation (BRAQUE) to a large bibliographic information retrieval system. The interface scheme is based upon a progressive development of the capabilities that the final interface is going to have; the framework for the interface is articulated on the basis of an information seeking strategy model (ISS), a cognitive task analysis (CTA) and a two-level hypertext model [5] for information systems. The design reported in the paper has been translated into a prototype of an operational system.

*Pollard* in [30] reports on work that provides an online thesaurus as an interface which helps the end-user in his information search; the thesaurus is presented to the user as a browsing interface implemented through a hypertext. Through this interface the user uses the information stored in a bibliographic database. The paper presents the design and implementation of the interface established using a commercially available hypertext system.

## 6.5 Evaluation of HIR systems

Information retrieval evaluation techniques like computing precision and recall are not directly usable in the evaluation of characteristics of hypertext information retrieval systems. It is therefore necessary to develop new procedures and tools that establish a relationship between present evaluation efforts to previous evaluation work in information retrieval, and at the same time be able to effectively evaluate new system capabilities.

In [20, 21] a proposed model is tested by carrying out two experiments that use a text document collection to relate the results to previous findings in the information retrieval area. Some insights into the development of evaluation techniques for hypermedia systems are given together with some more general results from a combination of query-based and browsing-based retrieval capabilities.

*Croft* and *Turtle* in [19] also deal with the problem of evaluating HIR systems. In fact a comparison of performance of the strategies used in two retrieval models is made; a probabilistic retrieval model incorporating inter-document links with strategies that ignore the links versus a heuristic spreading activation strategy. The findings show that a hypertext retrieval model based on inference networks can be considered just as effective as spreading activation.

# 7 NETWORKED HIR

Technical and technological changes have occurred in recent years because of the widespread use of the *World Wide Web (WWW)* technology developed in the framework of the WWW project [12]. These changes are having an important impact on HIR applications and because of that, general aims and underlying standards are now examined here.

The WWW project is a wide-area hypertext information retrieval initiative aiming at giving universal access to a large volume of documents over the Internet. There are WWW servers and clients, where the WWW servers are managing and making publicly available collections of hypermedia documents, and the clients incorporate a browser that permits the access to any WWW server and the managed hypermedia documents. Mosaic and NetScape are examples of WWW browsers. The adoption of one of these browsers makes the navigation on a hypermedia document easy, since each browser provides a "point-and-click" interface with all the WWW built-in functions for browsing through a hypermedia.

If one needs to make available a hypertext or hypermedia to end-users, it is very common to see the organisation taking the decision of making information available through the WWW. Many organisations have decided to make available their information in this way, because the end-user does not need to have specific supplied software on his platform (PC, Mac or X-terminal) to access the hypermedia as there are WWW browsers available for almost all platforms. It is thus sufficient that the user has one of the free and publicly available WWW browsers, and has access to the Internet site where the hypermedia is hosted. The use of different WWW information sources is spreading very fast over different categories of applications.

The success of the WWW project is based on three opportune and effective Internet standards [12, 13]:

- *Internet addresses*: the Internet method of addressing resources;
- *HTTP: HyperText Transfer Protocol*: a protocol for communicating hypertext documents on the Internet;
- *HTML: HyperText Markup Language*: a hypertext markup language developed in the context of the WWW project for marking documents; HTML enables the transformation of a flat text into a hypertext. HTML is an application of SGML [35].

The availability of both WWW client and server technology shifts the focus towards applications that can be developed and made available through WWW technology, instead of on the development of an interface, client and specific application at server level.

Experiences that have been gained with using networked IR on textual collections like *wais*, the wide area information service, are related to the use of hypertext systems for IR operations similar to the WWW experiment and its interfaces. In such applications, non-text media are used to support and explain a resulting search obtained from formulating and running a query based solely on indexing terms. Thus a combination of searching facilities based on text and other media would be of great help in the searching interaction between the end-user and the retrieval system.

A prototype for a system with those HIR capabilities has been presented in [7]. Other efforts are to be developed by other research groups and organisations. These initial results show the importance and difficulty of this problem and indicate the value of more in-depth study. The problem of the automatic authoring of *multimedia documents* remains to be addressed.

## Acknowledgements

This chapter is based on work in *Hypertext Information Retrieval* conducted over many years by the *Information Management Systems* research group of the Department of Electronics and Informatics of the University of Padua. I take this opportunity to thank all my colleagues in the group for the fruitful discussions we have had together on this topic.

## REFERENCES

[1] Aalbersberg, I.J. (1992). *Incremental Relevance Feedback.* In: N. Belkin, P. Ingwersen and A.M. Pejtersen (Eds.), Proc. 15th. SIGIR Conference, Copenaghen, (Denmark), 11–21.

[2] Agosti, M. (1988). *Is Hypertext a New Model of Information Retrieval ?* In: Proc. 12th International Online Information Meeting. Learned Information, Oxford, Vol. I, 57–62.

[3] Agosti, M. (1991). *New Potentiality of Hypertext Systems in Information Retrieval Operations.* In: H.-J. Bullinger (Ed). Human Aspects in Computing. Elsevier Science Publishers, Amsterdam, The Netherlands, 317–321.

[4] Agosti, M., Colotti, R. and Gradenigo, G. (1991). *A Two-Level Hypertext Retrieval Model for Legal Data.* In: A. Bookstein, Y. Chiaramella, G. Salton and V.V. Raghavan (Eds.), Proc. 14th ACM-SIGIR Conference, Chicago (USA), 316–325.

[5] Agosti, M., Gradenigo, G. and Marchetti, P.G. (1992). *A Hypertext Environment for Interacting with Large Textual Databases.* Information Processing and Management, 28(3), 371–387.

[6] Agosti, M. and Marchetti, P.G. (1992). *User Navigation in the IRS Conceptual Structure through a Semantic Association Function.* The Computer Journal, 35(3), 194–199.

[7] Agosti, M., Crestani, F. and Melucci, M. (1995). *Automatic Authoring and Construction of Hypermedia for Information Retrieval.* ACM Multimedia Systems, 3(1), 15–24.

[8] Agosti, M., Crestani, F. and Melucci, M. (1996). *Design and Implementation of a Tool for the Automatic Construction of Hypertexts for Information Retrieval.* Information Processing and Management, 32 (in print).

[9] Allan, J. (1995). *Relevance Feedback with Too Much Data.* In E.A. Fox, P. Ingwersen and R. Fidel (Eds.). Proc. 18th ACM-SIGIR Conference, Seattle, (USA), 337–343.

[10] Belkin, N.J. and Croft, W.B. (1987). *Retrieval Techniques.* In: M.E. Williams (Ed). Annual Review of Information Science and Technology (ARIST), 22, 109–145.

[11] Belkin, N.J., Marchetti, P.G., and Cool, C. (1993). *BRAQUE: Design of an Interface to Support User Interaction in Information Retrieval.* Information Processing and Management, 29(3), 325–344.

[12] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H.F. and Secret, A. (1994). *The World-Wide Web.* Communications of the ACM, 37(8), 76–82.

[13] Berners-Lee, T. and Connolly, D. (1995). *Hypertext Markup Language - 2.0.* HTML Working Group, Internet-draft, 22 September 1995.

[14] Bieber, M. and Isakowitz, T. (Guest Eds), (1995). *Introduction to the Special Issue on Designing Hypermedia Applications.* Communications of the ACM, 38(8), 26–29.

[15] Bruza, P.D. and van der Weide, T.P. (1992). *Stratified Hypermedia Structures for Information Disclosure.* The Computer Journal, 35(3), 208–220.

[16] Bush, V. (1945). *As We May Think.* Atlantic Monthly, 176, 101-108.

[17] Conklin, J. (1987). *Hypertext: an Introduction and Survey.* IEEE Computer, 20(9), 17–41.

[18] Croft, W.B. and Thompson, R.H. (1987). *I3R: a New Approach to the Design of Document Retrieval Systems.* Journal of the American Society for Information Science, 38(6), 389–404.

[19] Croft, W.C., and Turtle, H.R. (1993). *Retrieval Strategies for Hypertext.* Information Processing and Management, 29(3), 313–324.

[20] Dunlop, M. (1991). *Multimedia Information Retrieval.* PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, UK.

[21] Dunlop, M. and van Rijsbergen, C.J. (1993). *Hypermedia and Free Text Retrieval.* Information Processing and Management, 29(3), 287–298.

[22] Frisse, M.E. (1988). *Searching for Information in a Hypertext Medical Handbook.* Communications of the ACM, 31(7), 880–886.

[23] Frisse, M.E. and Cousins, S.B. (1989). *Information Retrieval from Hypertext: Update on the Dynamic Medical Handbook Project.* In Proc. Hypertext '89 Conference, Pittsburgh, (USA), 199–212.

[24] Furuta, R., Plaisant, C. and Schneiderman, B. (1989). *Automatically Transforming Regularly Structured Linear Documents into Hypertext.* Electronic Publishing, 4(2), 211–229.

[25] Lancaster, F.W. (1986). *Vocabulary Control for Information Retrieval (2nd Ed).* Information Resources, Arlington, Virginia, 1986.

[26] Lucarella, D. (1990). *A Model for Hypertext-based Information Retrieval.* In: Rizk, A., Streitz, N. and André, J. (Eds). Hypertext: Concepts, Systems, and Applications, Cambridge University Press, 81–94.

[27] Lucarella, D. and Zanzi, A. (1993). *Information Retrieval from Hypertext: An Approach Using Plausible Inference.* Information Processing and Management, 29(3), 299–312.

[28] Nanard, J. and Nanard, M. (1995). *Hypertext Design Environments and the Hypertext Design Process.* Communication of the ACM, 38(9), 49–56.

[29] Nielsen, J. (1990). *Hypertext and Hypermedia.* Academic Press, Boston.

[30] Pollard, R. (1993). *A Hypertext-based Thesaurus as a Subject Browsing Aid for Bibliographic Databases.* Information Processing and Management, 29(3), 345–357.

[31] Rada, R. (1992). *Converting a Textbook to Hypertext.* ACM Transactions on Information Systems, 10(3), 294–315.

[32] Salton, G. and Buckley, C. (1989). *Automatic Generation of Content Links for Hypertext.* Research Report, Department of Computer Science, Cornell University, Ithaca, New York, June 1989.

[33] Salton, G., Allan, J. and Buckley, C. (1993). *Approaches to Passage Retrieval in Full Text Information Systems.* In R. Khorfage, E. Rasmussen and P. Willett (Eds.), Proc. 16th ACM-SIGIR Conference, Pittsburgh, (USA), 49–58.

[34] Salton, G., Allan, J. and Buckley, C. (1994). *Automatic Structuring and Retrieval of Large Text Files.* Communications of ACM, 37(2), 97–108.

[35] SGML (1986). *ISO Standard Generalized Markup Language.* ISO 8879: 1986.

[36] Shneiderman, B. and Kearsley, G. (1989). *Hypertext Hands-on! An Introduction to a New Way of Organizing and Accessing Information.* Addison-Wesley, Reading, MA.

[37] Smeaton, A.F. (1995). *Building Hypertext under the Influence of Topology Metrics.* In Proc. International Workshop on Hypermedia Design (IWHD), Montpellier, (France).

[38] Smith, J.B. and Weiss, S.F. (1988). *An Overview of Hypertext.* Communications of the ACM, 31(7), 816–819.

[39] van Rijsbergen, C.J. (1979). *Information Retrieval (2nd Ed).* London, Butterworths.

[40] Yankelovich, N., Haan, N.K., Meyrowitz, B.J. and Drucker, S.M. (1988). *Intermedia: the Concept and the Construction of a Seamless Information Environment.* IEEE Computer, 21(1), 81–96.