Pergamon

Information Processing & Management, Vol. 33, No. 2, pp. 133–144, 1997

0306-4573/97 $17+0.00

PII: S0306-4573(96)00058-1

ON THE USE OF INFORMATION RETRIEVAL TECHNIQUES FOR THE AUTOMATIC CONSTRUCTION OF HYPERTEXT

MARISTELLA AGOSTI, FABIO CRESTANI and MASSIMO MELUCCI
Dipartimento di Elettronica e Informatica, Università di Padova, Italy

Abstract—The first part of the paper briefly introduces what automatic authoring of a hypertext for information retrieval means. The most difficult part of the automatic construction of a hypertext is the creation of links connecting documents or document fragments that are semantically related. Because of this, to many researchers it seemed natural to use IR techniques for this purpose, since IR has always dealt with the construction of relationships between objects mutually relevant. The second part of the paper presents a survey of some of attempts toward the automatic construction of hypertexts for information retrieval. This part will identify and compare scope, advantages and limitations of different approaches. The aim of this survey is to point out the main and most successful current lines of research. © 1997 Elsevier Science Ltd

1. AUTHORING AND CONSTRUCTION OF HYPERTEXT

Any *hypertext* can be built by its author or groups of authors from scratch. However, nowadays the most common situation is the construction of a hypertext starting from a "flat" collection of "flat" documents available in a digitalized form. Building the hypertext requires then three main steps:

1. **Design**: identification and design of the target hypertext application; that is, what kind of hypertext application the author wants to produce given the target final user community.
2. **Authoring and Construction**: transformation of the initial single big document or collection of documents into a hypertext; the process requires the initial input to be fragmented and links between fragments to be built.
3. **Publishing**: making the hypertext available to the potential user community. To reach this final aim it is necessary to have a presentation and browsing tool that makes available to final users human–computer interaction (HCI) capabilities. This phase is heavily dependent on the available technology at the client/server and network levels.

Authoring, the phase we are most interested in, requires (1) the identification of the fragments of the original complete document(s) that will constitute the nodes of the hypertext, and (2) the creation of all the necessary links among nodes. The authoring process can range from a completely manual to a completely automatic one. A *design methodology* and a *conceptual reference architecture* are needed to support the authoring or construction process. At present, it is common practice to manually author a hypertext. However, if the initial collection of documents is of large proportions and/or also consists of multimedia documents, a completely manual authoring can be impossible to achieve. It is therefore important to have automatic techniques for the segmentation of documents, tools for an automatic generation of links, and procedures for the automatic updating of the hypertext to insert, modify, and cancel part of it over time.

The paper focuses on *Information Retrieval* (*IR*) techniques for the automatic authoring of a hypertext. The reason why we chose to use IR techniques for the automatic authoring of a hypertext lies in the fact that the IR area deals with methods and techniques for *content-based* management and retrieval of information. Since the most difficult part of the automatic construction of a hypertext is the building up of links that connect semantically related

documents or document fragments, it is natural to concentrate on IR techniques, that have always dealt with the construction of relationships dependent on the mutual relevance of objects to relate.

It should be noticed that we concentrate on *hypertext*, because *hypermedia* and *multimedia* automatic authoring methods are still in their infancy, together with being heavily dependent on present day technology. On the contrary, the area of hypertext construction and authoring has been heavily addressed in the recent past, and some automatic construction tools are now available.

2. A TYPOLOGY OF HYPERTEXT LINKS

In this Section we would like to draw the reader's attention to the different types of links that can be built in a hypertext. It is by making use of the different characteristics of these different types of links that the hypertext can be built.

In a hypertext, a link implements a logical connection between two related nodes:

- the *origin node*: the node from which the connection starts;
- the *destination node*: the node where the connection 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 the hypertext. Most hypertexts are built making use of the following three types of links:

- *Structural links*: connect nodes of the hypertext that are related by the structure of the document itself. Examples of this type of link are links connecting a chapter with the chapter that follows it, or a table of contents with each section reported in it. If the transformation from a flat document that is a book to hypertext is made, all the chapter/section/subsection structural connections can be rendered in the hypertext using structural links. Each structure (e.g. tree, graph) can be rendered using structural links.
- *Referential links*: are based on some sort of reference the author of the original document has used. A typical link of this sort is the link that implements a reference between the source document and a document that is cited by it.
- *Associative links*: represent undefined associative connections between nodes. They are built making use of content-based connections between fragments of text of the same document or documents of the same collection.

All those links are made explicitly available to the user through the hypertext network. Another type of link, the *aggregate link*, is made available only in few hypertext systems, where the aggregation abstraction mechanism is designed and implemented to give the hypertext designer the possibility of aggregating nodes that together form a new kind of node; this mechanism has been initially introduced and made available in semantic data models (Schiel, 1989) but it is still seldom available in operative database management and hypertext systems (Smith & Smith, 1977).

Another classification of links is very useful to add, that of explicit/implicit link:

- 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 in a node. An implicit link can be activated using a word present in a node. For example, if a user asks to see all nodes that contain a particular word, all nodes containing that word can be made available to the user, and these implicit links are created at run time. This means that an implicit link does not link a pair of nodes, but it is implicitly present in the node text and it is created and made available at run time. An example of this process is described in (Aalbersberg, 1992).

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* the hypertext. In order to navigate the hypertext, the user needs a tool able to follow the links: this capability is usually provided by a tool called a *browser.* A browser usually incorporates both navigation and browsing facilities. 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 it is by searching the network for some word, string, keyword or attribute value which nodes have to share, and 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. Most present hypertext browsing tools of stand alone hypertext systems cannot usually provide exact match retrieval techniques that use a query language based on Boolean logic, and that are available in most of the present operational information retrieval systems. On the contrary, the World Wide Web (WWW) is a large hypertext of fairly sophisticated search engines, though they are only vaguely incorporated into the browsers themselves.

The capability of easily linking different pieces of information, 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 feature of hypertext systems together with other features relevant to information retrieval operations, work has started in the new area of Hypertext Information Retrieval (HIR). Some of the initial papers addressing the issues related to the combination of hypertext and information retrieval capabilities to produce a new sort of innovative information management tools are Agosti (1988, 1991), Croft & R. H. Thompson (1987), the collection of papers in Agosti (1993) presents the research results at the time of publication of that special issue, and Agosti & Smeaton (1996) have made the effort of presenting in an integrated way the merging of the two areas of IR and hypertext.

3. A SELECTION OF EXPERIENCES IN THE AUTOMATIC CONSTRUCTION OF HYPERTEXT

For reasons outlined in the previous Sections, IR has often provided valuable techniques for automatic or semi-automatic hypertext construction. These techniques have been used for many different purposes, going from the simple indication of interesting links to the hypertext author, to the evaluation of the effectiveness of the resulting hypertext. The increasing importance of having techniques for the fully automatic construction of hypertexts has motivated many research groups in the search of effective IR techniques for this purpose. In this Section we will survey some of these attempts, trying to point out their scope, advantages and limitations. This survey is not trying to be comprehensive of all the approaches attempted in this direction. We are aware that some attempt will slip through our survey, however the main purpose of this paper is just to point out the main and most successful current lines of research.

Our survey will report past and current experience organized around the "schools" (universities and research institutions) where this kind of research was conducted and is still going on.

3.1. Cornell University

It is natural to start our survey by reporting the long experience of work done in the direction of the automatic construction of hypertexts at Cornell University under the direction of Professor Salton. The longstanding experience in statistical based approaches to IR achieved by this research group provides the background from which their techniques for the automatic

construction of hypertext come from.

Salton & Buckley (1992) do not directly tackle the problem of automatic construction of a hypertext, but they propose a technique that can be used to create links between text segments, that practically builds up a hypertext at retrieval time. The technique proposed here is the first attempt to use vector similarity to produce a network of text segments that are semantically related.

The basic idea is to use the normalized $tf \cdot idf$ weighting schema in the context of the vector space model to evaluate the similarity between two text segments. The technique is used with two different scopes:

- if the text segment corresponds to the whole document then the technique can be used to produce a global measure of similarity between two documents or a document and a query (global similarity);
- if the text segment is a paragraph or a sentence inside a document then the technique can be used to produce a similarity between documents that is based on the their maximum pairwise similarity between text segments (local similarity).

The authors show that there is a strict correlation between global and local similarities, and that local similarity is more precision oriented that global similarity and can be used to refine the retrieval result of the latter.

The same technique that they used to retrieve text segments in response to a query can be used iteratively to link text segments at retrieval time. This technique works in the following manner:

1. retrieve, in response to a query, a set of m text segments using global similarity;
2. refine the retrieved set by rejecting all but k text segments. This can also be done in two ways: by setting the value of k first, or by employing a local similarity threshold and accept the k segments that are over that threshold. In the second case the value of k cannot be controlled directly;
3. use the retrieved set of k segments as a new set of queries and for each of them restart the process. The process can be repeated n times. At each iteration link the k segments (queries) with each of the k new text segments accepted in response of each query.

The process produces a tree where nodes can reappear at different levels, and whose depth and breadth can be controlled by carefully choosing respectively m, n and k.

The major advantage of this approach is that it can be performed at retrieval time. This could appear as a disadvantage, since it could take a long time to produce a large number of links, but links can then be stored and used subsequently when one of these $n \cdot k$ text segments is retrieved in response to a new query. However the process does not take into consideration the problem of updating the link structure once new documents are added to the collection. Moreover it is not clear if the text segments traversal can only be performed going down from the root to the leaves of the tree or also in the opposite direction. Also the problem of the identification of the correct dimension of the text segments is left unsolved.

As usual, Salton follows an experimental approach, carefully substantiating with experimental results the value of each step of the technique. The paper is clear and the technique proposed is well explained using numerical examples.

Salton & Allan (1994) take the previous technique of combining global and local similarity measures another step further. A graph representation is introduced that resembles very much a hypertext. Nodes of this graph can be documents or textual fragments (paragraphs, sentences) extracted from documents. When nodes represent documents, a global measure of similarity is used to measure closeness among documents. This is based on the inner product of the weighted vectors representing the documents in the Vector Space Model. In this case, term weights are computed using the normalized $tf \cdot idf$ weighing scheme. Similarly, when nodes represent text segments, the same inner product is used to measure their pairwise similarity. However, in this case non-normalized term weights are used, to give preference to longer matching sentences that are more indicative of coincidences in text meaning. An accurate analysis of the structure of a document can be obtained by putting the nodes representing text fragments along a circle and drawing a line whenever the pairwise similarity is over a predetermined threshold. Using this

technique it is possible to decompose documents by identifying homogeneous parts (sets of text segments) of the document. The same technique could be used to link parts of the document that have strong relationships between them.

Local similarity is proposed in this paper as a precision filter, that can be used to discard documents that may have a high global similarity with the query due to language ambiguities, but that have a low local similarity with the query. This technique can be used also to perform hypertext automatic authoring, as a tool for automatic document decomposition and structuring.

In Salton *et al.* (1994) a range of different IR text manipulation methods and strategies is addressed. This paper is probably the final and most widely known of the above series of papers on the use of a global–local matching technique for text retrieval and text structuring. The paper is nicely written with lots of clear examples and with a clear introduction to what is the purpose of their research. The technique they propose, an extension of the vector space model to include a precision oriented device (a local matching), is the same already presented in Salton & Buckley (1992) and Salton & Allan (1994) but here it is addressed with the dual purpose of retrieval and text structuring. It is this second use that is most interesting to us, since this technique can be used for the automatic construction of a hypertext at retrieval time. The use of the global–local matching technique already explained above assures that the links are semantically oriented. This paper reports a nice classification of the different techniques that can be used for automatic text structuring. According to Salton there are three classes:

1. "breadth m–depth n" search: fixed number of documents accepted in response to a query and fixed number of iterative searches;
2. decreasing levels of similarity: the number of accepted documents is variable since it is determined by a similarity threshold, the threshold is then progressively increased to produce a self contained map;
3. clustering: performed on the results of a search and incorporated into the final ranking.

Examples of the practical use of these technique are reported in the cited paper.

To conclude this long section, the work at Cornell has been very successful, so much that it was followed by many other schools. However, some questions that Salton pointed out as important at the very beginning of his experience with hypertext automatic construction (Salton & Buckley, 1989) seem to remain partially unsolved:

1. The resulting hypertext should be tested to see if it is useful for the IR purposes, both from a system and user's point of view. Although good from an IR point of view such hypertext might not be good for user browsing.
2. Is this technique effective for documents covering heterogeneous subjects? The problem requires further attention, but it seems possible to anticipate that the effectiveness of a technique for the automatic construction of hypertexts depends on the extent of the subjects of the documents. Experiments for linking related texts in a large collection of unrestricted subject matter have been conducted at Cornell by Salton & Allan (1993) and Salton *et al.* (1996). The results of these experiments need to be completed with formal evaluation data before being able to completely assess the success of the proposed techniques.
3. The goodness of the resulting hypertext is related to the good tuning of some parameters, such as the similarity threshold. Is there an automatic technique for setting these parameters? How do they affect the effectiveness of the hypertext?

These questions are still at the very heart of the research on the automatic construction of hypertexts.

3.2. University of Massachusetts

The previous Section has addressed the work at Cornell; James Allan has recently moved at the University of Massachusetts, but he remains active in the area with new results. In particular he continues the work of his PhD Thesis (Allan, 1995), addressing the problem of discovering link types. The technique he proposed provides a way of setting up links between passages of

documents. The novelty of this work is in the use of classical IR techniques to determine the type of relationships incurring in a hypertext, where nodes represent topics. Allan also addressed the problem of the number of links.

To reduce the number of links and to make easier the visualization of the resulting graph, some techniques for merging are suggested and described. The techniques proposed by Allan for the automatic link type identification are based on some values calculated on the merged links. For example, to identify a summary link, we can compute the amount of unlinked text that was added to a link endpoint during the link merging.

The author points out a few directions that should be followed by further research in automatic authoring based on IR techniques. Among them, additional work should be done with regards to heterogeneous documents, or documents that have been written with a poorly regular writing style; most proposed techniques for automatic authoring are based on the assumption that documents are quite well-segmented into passages. Some recent research findings are presented by Allan in a paper of this special issue (Allan, 1997).

3.3. University of Liverpool

Rada (1992) addresses the combination of structural links and content links. The author was the first to distinguish between first-order and second-order hypertext. First-order hypertexts use only structural links based on the document mark-up and determined by the document author. Structural links of this kind include links connecting outline headings, citations, cross-references and indices. In second-order hypertexts, links are not explicitly put in the text by the author, but are detected using some automatic procedures. This distinction is still in use and it is used in some manner by everybody in the field. The use of first and second order links in the same hypertext enables it to reflect both the structural schema of the source documents (the document author's schema), and an alternative schema, reflecting the way index terms are distributed across the documents. 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.

In the cited paper, second-order links were set up between index terms using co-occurrence data. The technique is therefore quite simple since it is simply based on the use of a threshold on the co-occurrence data to set up or not links between index terms. The data used to test the technique was a textbook on hypertexts that was originally written as a first-order hypertext.

The main contribution of this work is not in the single technique used, but in showing how to develop an integrated methodology for transforming a book into an hypertext. This methodology was tested using four tools: Guide, HyperTies, Emacs-Info and SuperBook.

3.4. University of Maryland

One of the earliest work in the field of the automatic transformation of text into hypertext was performed by Furuta *et al.* and is reported in Furuta *et al.* (1989a). This paper illustrates the design and 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 cards or manual pages. The resulting hypertext is made of nodes corresponding to the document parts connected by means of structural links. The methodology 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 organization 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, the collection of dissertation abstracts they used for their experiments. Larger, or less regular and consistent documents, for example

scientific papers, would require some manual intervention to catch content-based links, that is, links non-explicitly inserted in the documents and that are meant to represent semantic "aboutness". This conclusion is supported by the results of an extensive experimentation of the proposed methodology that is reported in Furuta *et al.* (1989b). Unfortunately, it is the latter kind of links that are the most interesting from an IR point of view, since we are mainly interested in semantic navigation and browsing of the document collection.

3.5. NEC Corporation

NEC, in collaboration with the University of Maryland, addressed the problem of identifying aggregates within existing hypertexts mostly by analysing link patterns. The aggregate identification method, proposed by Botafogo *et al.* (1992) and Botafogo (1993), is pertinent to this survey because it can be used as a guideline to design algorithms for the automatic construction of hypertexts. The authors address a well-known problem within hypertext called "user disorientation". User disorientation comes from the high number of links we need to follow to get interesting nodes. A high number of links sometimes indicates a too complex hypertext structure. The hierarchical structure is often the most natural organization of information. Hypertext authors often start to create hypertexts in a hierarchical way, but this guideline is often lost because of the intrinsic network nature of cross-referenced documents.

The authors propose an improvement of authoring techniques through the identification of specific hypertext sub-structures, such as clusters or hierarchies, and the definition of metrics called "compactness" and "stratum". These tools are aimed to help authors in writing hypertexts, and then to solve the user disorientation problem. Hierarchies are important because the root is a node that can reach all the other nodes with a low number of links. Clusters are sets of related nodes which are identified by analysing the hypertext structure. Compactness measure helps the hypertext author to assess the density of linking, i.e. roughly speaking the mean number of links per node. Stratum measure gives an account about the number of links to be traversed to get a node starting from an another. By finding the roots or clusters of a hypertext, or by adopting some metrics, authors can have a more clear idea about structure, and then they can indirectly evaluate the effectiveness of the hypertext.

The main conclusion authors draw is that one needs some guidelines to keep hypertext construction in control especially whenever scaling up is required. This is essential if automatic construction processes have to be performed, as illustrated in Section 3.6. Algorithms used to find hierarchies or clusters of hypertexts do not scale up well when very large hypertexts are analysed. Anyway, if one assume of performing hierarchy or cluster extraction in an incremental manner, the computational problems can be overcome by using some modified algorithm. For example, a modified version of the breadth-first-search algorithm for hierarchy identification can be performed only when new documents are authored. Even though hierarchies, clusters, and metrics are very useful, one cannot forget that a good evaluation of hypertext does not come only from objective numerical sources, but mainly both from the author experience and from tests with real users. As authors claimed in Botafogo *et al.* (1992), it is necessary to verify whether, and to assess the degree to which, employed metrics or strategy for aggregate identification work for diverse link types as well: it is reasonable to think that different metrics need to be devised for other hypertext properties, such as "readability" or "comprehensibility".

3.6. Dublin City University

The approach reported in Smeaton & Morrissey (1995) is extremely interesting because it addresses "head on" the problem of the automatic construction of a hypertext from a text using techniques from IR that are dynamically guided by a overall measure of how good the resulting hypertext is. The approach is not new from the point of view of the techniques used: the classical $tf \cdot idt$ measure of node–node similarity is used, while the "goodness" of the resulting hypertext is measured using the Botafogo measure of compactness (see Section 3.5). The novelty of the approach proposed is in the use of the Botafogo's measure as a dynamic control structure that provides a cut off point in the incremental addition of links in the hypertext. The resulting

approach provides a technique for the selective creation of links that is based on a measure of the overall topology of the hypertext that controls the adding of a new link in relation to its influence on the overall hypertext topology.

Furthermore, Smeaton (1995) analyses the use of the compactness measure proposed by Botafogo, for the construction of hypertexts. The analysis is performed by evaluating the compactness measure using a web robot on four different hypertexts. These hypertexts are based on four quite different topologies and are either manually (two of them) or automatically constructed (the other two), with links reflecting only the structure of the document (first-order hypertexts) and/or links reflecting content similarity between nodes (second-order hypertexts). The effect on the compactness measure of the different topologies is analysed and some interesting conclusions are drawn. The major conclusion, however, is that the compactness measure is not particularly useful in guiding the automatic construction of large hypertexts, but it could be useful (experiments in Dublin are under way in testing this) for the automatic authoring of sub-parts of the hypertext.

The major contribution of these works 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 authors highlighted, some very compact hypertexts 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 the browsing by providing more information about links, such as, for example, information about the link types.

3.7. Brown University

The work carried out by Coombs (1990) aims at providing an integrated environment for browsing, searching, and automatic linking of documents in IRIS Intermedia. The integration of such different capabilities comes from the need for breaking through the classical hierarchical document structure given by a file system that imposes a single path to access the desired information. This approach shares the same target of most research work in the area, that is, to provide final users and authors with tools and structures which reduce the risk of getting lost in the hypertext due to the complex hypertextual topology.

Intermedia users manually create links to re-find information by enabling them to create trails through information that they have found valuable. Users can then superimpose many different perspectives at once on a single document set corresponding to the various personal ways of viewing information. The architecture provided for search readily supports automatic linking of documents and passages. The Intermedia AutoLink function assists users in initiating a "web" on a topic by linking, for example, all references to a subject into an overview document. AutoLink can also support users in resolving coded cross references. For example, numerical references can be replaced by a more explicit and clear string representing the linked topic. The hypertext author can then perform much of the work automatically instead of working through the thousands of articles by hand.

The automatic linking capability provided within Intermedia can be used as an effective tool for authoring. This initiative is within an educational project carried out at Brown University. Intermedia approach helps to keep users in control by supporting them in creating document organizations. Such organizations have an educational value that cannot be achieved by automatic authoring. Students who manually create links retain knowledge longer than those merely receive and memorize. However a support for linking does not completely solve the problem due to the size of the document set: it is the size and the number of documents that make automatic linking a hard task.

3.8. Padua University

Agosti & Crestani (1993) proposed a design methodology for automatically constructing an IR hypertext by putting together various well established IR techniques of linking IR objects. This methodology, which will be briefly described in the following, is based on a modified

version of the EXPLICIT conceptual model (Agosti *et al.*, 1991).

The aim of the methodology is to enable users of large document collections to browse the document base in a natural way, navigating through connections representing statistical or semantic relationships between documents.

It has been long recognized that the complexity of data modelling in IR is mainly due to the complex nature of the relationships between the different IR objects: documents and auxiliary data. (Auxiliary data are all those objects that are used to index the documents, e.g. index terms, classification structures, and thesauri.) Therefore the proposed design methodology is based on a conceptual architecture that is structured on three levels: *document level*, which contains the collection of documents; *index term level*, which contains the index terms that are linked to the documents they are meant to index; *concept level*, which contains sets or classes of index terms, where a concept is a higher level object than an index term, so a concept can be connected to several index terms that are meant to represent it.

The initial input is the usual set of IR raw data: a flat large document collection. Documents are available as individual unrelated objects. Then the documents are indexed, constructing the collection of index terms that are connected to documents through the automatic indexing process. It is by means of this process that individual or groups of terms found in documents become index terms, assuming a representational power that places them on a higher level of abstraction than the documents. The conceptual level can be built if a set of concepts representing the application domain of interest can be found; if the available tool is a thesaurus, the structure of its associations can be directly mapped into a network structure: concepts are mapped to nodes and concept relationships to links. If this tool is not available, it becomes necessary to build up the network of concepts manually. The first essential step is to identify a set of concepts and their relationships. Afterwards the semantic associations between index terms and concepts can be built using different formal approaches. The approach described in Agosti & Marchetti (1992) and named "semantic association" permits the automatic construction of links between index terms and concepts. Afterwards a technique for finding relationships between index terms using information provided by statistical analysis of index term occurrence in documents is adopted. For the automatic set up of links between documents, statistical techniques are used very similar to those employed for the construction of links between index terms. It is worth considering at this point that most operational IR systems use only the document–index term links in the retrieval process, and only very few operational IR systems enable the user to take advantage of relationships like those established by concept–concept and index term–concept links, and they are used only as an aid to query formulation. Relationships like those constructed with this methodology and represented by index term–index term and document–document links are used only in few experimental IR systems.

Using that methodology for the automatic construction of hypertexts as the methodological reference framework, the tool TACHIR (a Tool for the Automatic Construction of Hypertexts for IR) has been designed and implemented to be able to automatically construct a hypertext from a flat document collection. The structure of the built hypertext reflects the three level reference model. The hypertext is implemented using the HTML hypertext mark-up language, the mark-up language of the World Wide Web project. It can be distributed on different sites and different machines over the Internet, and it can be navigated using any of the interfaces developed in the framework of the World Wide Web project, such as for example NetScape (Agosti *et al.*, 1994, 1995).

At present, ways of improving user access to the hypertext nodes by adding techniques to *query* the hypertext are under study. This would enable the user to directly access nodes that are good starting points for browsing and from which he could reach relevant nodes more quickly. And techniques for using the hypertext to perform retrieval of relevant nodes by "constrained spreading activation" (Crestani, 1996) are also under investigation with the aim of spreading the user relevance assessment to nodes that are related to them with some degree of significance. The result would be a ranking of a large set of nodes of the IR hypertext that will be presented to the user using an appropriate interface. The user could then assess again the retrieved set and produce a new spreading of activation, thus making the retrieval process an iterative and interactive process.

3.9. Other works

Research on the automatic construction of hypertexts is not only going on at universities and research institutions. Some commercial companies are investing in this field that represent one of the future lines of exploitation of multimedia documentation (we have seen already the example of NEC Corporation). We therefore believed that we should mention at least a few important products that were released recently into the market.

3.9.1. Productivity Edge by PRC Inc. Recently we came across a report on a new product described as a "document management product" called Productivity Edge, by PRC Inc., a commercial company that has strong links with the group of E. Fox and the Virginia Tech, U.S.A.

The product is based on a incremental clustering technique proposed by F. Can, that enable the semi-automatic generation of links between sentences, paragraphs, or entire documents. The link generation is performed at retrieval time and is user directed. The speed of the technique is assured by the use of the Fulcrum retrieval engine.

One of the major advantages of this product is the fact that the hypertext produced is written on the fly using HTML and therefore can be browsed using any WWW browser. The product is currently being evaluated. The quality and quantity of the links generated is going to be compared with that of a manually constructed hypertext.

3.9.2. Hyperties by Cognetics Corp. This system is a commercial version of the automatic authoring system developed by Furuta and whose details are reported in Furuta *et al.* (1989a). We will not enter into the details of this system here, since we already talked about it in Section 3.4.

3.9.3. LaTeX2HTML by Nikos Drakos. The LaTeX2HTML converter is a non-commercial system to translate LaTeX (Lamport, 1986) files into a web of HTML files; it has been developed by Drakos at The University of Leeds (Drakos, 1994).

The aim of the LaTeX2HTML project is to help authors to represent their textual material as a network of linked nodes. The most used documents for being translated into a hypertext by LaTeX2HTML are books, manuals, articles, or reports, which are well-structured documents. The rich hypertextual structure of these well-structured documents is available thanks to the fact that authors have already inserted anchors between documents or parts of them. LaTeX2HTML helps create a HTML hypertext by automatically recognizing implicit links, such as titles, and explicit links, such as references.

The most important features of LaTeX2HTML is the flexibility in specifying the desired node granularity, and the ability in managing mathematical formulas. The most significant limitations of LaTeX2HTML are due to the impossibility of recognizing all the LaTeX "dialects" that are defined by different writers through the definition of new commands, and to the lack in recognizing semantic links other than structural links.

4. THE EVALUATION ISSUE

The evaluation of automatically constructed hypertext IR (HIR) applications and systems is a very complex issue, because evaluation techniques used in IR (like precision and recall) are only partially usable in the evaluation of characteristics of hypertext information retrieval systems. It is therefore necessary to develop new methods that establish a relationship between present evaluation efforts to previous evaluation work in information retrieval. The method, at the same time, should in general evaluate effectively new system capabilities, and in particular address the evaluation of those applications and systems that have been automatically developed and made available to the users.

Some efforts have been reported for the evaluation of HIR applications. In particular, Dunlop (1991) and Dunlop & van Rijsbergen (1993) have tested a proposed model of hypermedia IR by carrying out two experiments that make use of 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 & Turtle (1993) 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.

A recent work that is relevant to the evaluation of automatic constructed hypertexts for IR is reported in Furner *et al.* (1996). This work reports on an experimental study of comparison of hypertexts, where different people were asked to produce a hypertext representation of the same full-text document. The investigation has required the calculation of measures of similarity between pairs of manually produced hypertexts. This has been achieved by representing the hypertexts as labelled graphs, in which the vertices and the edges of a graph were used to define the paragraphs and the inter-paragraph links, respectively, of a hypertext, and by then comparing the resulting graphs. Many different similarity measures are available for the comparison of such graphs but the extended experiments conducted by the authors suggest that they give broadly comparable results: specifically, all of the measures yield a very low degree of similarity between the various hypertext versions of each of the five full-text documents that were used, with several of the measures being monotonic with each other. The authors accordingly conclude that the structure of a hypertext document is crucially dependent upon the person who has created the links. This important experimental result implicitly supports the study and development of methods for the automatic authoring of hypertexts, since the application of a well founded automatic authoring method can produce a hypertext that does not incorporate only a specific designing and development view, but also it is reproducible starting from the same initial collection of flat documents.

The paper Blustein *et al.* (1997), of this special issue, addresses the problem of defining and making available methods for the evaluation of the quality of automatically constructed links.

5. CONCLUSIONS

In this paper we presented a survey of the use of IR techniques for the automatic construction of hypertext. Many are proposed approaches and it is quite difficult to distinguish between the successful ones and the unsuccessful ones. Until a well established evaluation methodology is available, like the one long in used in IR, it will always be very difficult to progress in this research area.

Acknowledgements—We would like to thank all the people that helped us, and in particular James Allan, who gave us many valuable suggestions. This work has been partially supported by the 1994 MURST 40% project on "Advanced Database Systems".

REFERENCES

Aalbersberg, I. (1992). Incremental relevance feedback. In *sigir*, pp. 11–22. Copenhagen, Denmark.

Agosti, M. (1988). Is hypertext a new model of information retrieval? In *Proceedings of the 12th International Online Information Meeting*, volume 1, pp. 57–62, Oxford. Learned Information.

Agosti, M. (1991). New potentiality of hypertext systems in information retrieval operations. In Bullinger, H., editor, *Human Aspects in Computing*, pp. 317–321. Elsevier Science Publishers, Amsterdam.

Agosti, M. (1993). Special issue on hypertext and information retrieval. *Information Processing & Management*, 29(3).

Agosti, M., Colotti, R., & Gradenigo, G. (1991). A two-level hypertext retrieval model for legal data. In *Proceedings of ACM SIGIR*, pp. 316–325, Chicago, U.S.A.

Agosti, M. & Crestani, F. (1993). A methodology for the automatic construction of a Hypertext for Information Retrieval. In *Proceedings of the ACM Symposium on Applied Computing*, pp. 745–753, Indianapolis, U.S.A.

Agosti, M., & Marchetti, P. (1992). User navigation in the IRS conceptual structure through a semantic association function. *The Computer Journal*, *35*(3), 194–199.

Agosti, M., Melucci, M., & Crestani, F. (1994). TACHIR: a tool for the automatic construction of hypertexts for

Information Retrieval. In *Proceedings of the RIAO Conference: Intelligent Text and Image handling*, pp. 338–357, Rockfeller University, New York, U.S.A.

Agosti, M., Melucci, M., & Crestani, F. (1995). Automatic authoring and construction of hypertext for Information Retrieval. *ACM Multimedia Systems*, *3*(1), 15–24.

Agosti, M. & Smeaton, A., eds (1996). *Information Retrieval and Hypertext*. Kluwer Academic Publishers, Boston, U.S.A.

Allan, J. (1995). *Automatic Hypertext Construction*. PhD Thesis, Department of Computer Science, Cornell University, Ithaca, NY, U.S.A. Available as Technical Report.

Allan, J. (1997). Building hypertext using information retrieval. *Information Processing & Management*, *33*(2), 145–159.

Blustein, J., Webber, R., & Tague-Sutcliffe, J. (1997). Methods for evaluating the quality of hypertext links. Submitted to Information Processing and Management, special issue on Information Retrieval and the Automatic Construction of Hypertext.

Botafogo, R. (1993). Cluster analysis for hypertext systems. In *Proceedings of ACM SIGIR*, pp. 116–125, Pittsburgh, PA, U.S.A.

Botafogo, R., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertextx: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, *10*(2), 142–180.

Coombs, J. (1990). Hyptertext, full text, and automatic linking. In *Proceedings of ACM SIGIR*, pp. 83–90, Brussels, Belgium.

Crestani, F. (1996). Applications of spreading activation techniques in information retrieval. *AI Review*. (In print).

Croft, W., & Thompson, R. H. (1987). I^3 *R*: a new approach to the design of Document Retrieval Systems. *Journal of the American Society for Information Science*, *38*(6), 389–404.

Croft, W., & Turtle, H. (1993). Retrieval strategies for hypertext. *Information Processing & Management*, *29*(3), 313–324.

Drakos, N. (1994). From text to hypertext: a post-hoc rationalisation of LaTeX2HTML. *Computer Networks and ISDN Systems*, *27*, 215–224.

Dunlop, M. (1991). *Multimedia Information Retrieval*. PhD Thesis, Department of Computing Science, University of Glasgow, Glasgow, U.K.

Dunlop, M., & van Rijsbergen, C. (1993). Hypermedia and free text retrieval. *Information Processing & Management*, *29*(3), 287–298.

Furner, J., Ellis, D., & Willet, P. (1996). The representation and comparison of hypertext structures using graphs. In Agosti, M. and Smeaton, A., eds, *Information Retrieval and Hypertext*, pp. 75–96. Kluwer Academic Publishers, Boston, MA, U.S.A.

Furuta, R., Plaisant, C., & Shneiderman, B. (1989). Automatically transforming regularly structured linear documents into hypertext. *Electronic Publishing*, *2*(4), 211–229.

Furuta, R., Plaisant, C., & Shneiderman, B. (1989). A spectrum of automatic hypertext construtions. *Hypermedia*, *1*(2), 179–195.

Lamport, L. (1986). LATEX *user's guide & reference manual*. Addison-Wesley Publishing Company, Inc., New York.

Rada, R. (1992). Converting a textbook to hypertext. *ACM Transactions on Information Systems*, *10*(3), 294–315.

Salton, G., & Allan, J. (1993). Selective text utilization and text traversal. In *Proceedings of ACM Hypertext'93*.

Salton, G., & Allan, J. (1994). Automatic text decomposition and structuring. In *Proceedings of the RIAO Conference: Intelligent Text and Image handling*, volume 1, pp. 6–20. Rockefeller University, New York, USA.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communication of the ACM*, *37*(2), 97–108.

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1996). Automatic analysis, theme generation, and summarization of machine-readable texts. In Agisti, M. and Smeaton, A., eds, *Information Retrieval and Hypertext*, chapter 3, pp. 51–73. Kluwer Academic Publishers, Dordrecht, NL.

Salton, G., & Buckley, C. (1989). Automatic generation of content links for hypertext. Thechnical report, Department of Computer Science, Cornell University, Ithaca, NY, U.S.A.

Salton, G. and Buckley, C. (1992). Automatic text struturing experiments. In Jacobs, P., ed., *Text-based intelligent systes: current research and practice in Information Extraction and Retrieval*, pp. 199–210. Lawrence Erlbaum Associates, Hillsdale, New Jersey, U.S.A.

Schiel, U. (1989). Abstraction in semantic networks: axiom schemata for generalization, aggregation and grouping. *SIGART Newsletters*, *107*, 25–26.

Smeaton, A. (1995). Building hypertexts under the influence of topology metrics. In *IWHD '95, International Workshop on Hypermedia Design*, Montpellier, France.

Smeaton, A., & Morrissey, P. (1995). Experiments on the automatic construction of hypertext from text. Technical Report CA-0295, School of Computer Application, Dublin, Ireland.

Smith, J., & Smith, D. (1977). Database abstractions: aggregation. *Communication of the ACM*, *20*(6), 405–413.