Information Retrieval 1, 91–114 (1999)

An Evaluation of Automatically Constructed Hypertexts for Information Retrieval

MASSIMO MELUCCI melo@dei.unipd.it
Dipartimento di Elettronica e Informatica, Via Gradenigo n. 6/A, 35131, Padova, Italy

Received January 9, 1998; Revised October 8, 1998; Accepted October 8, 1998

Abstract. This paper assesses the retrieval effectiveness of automatically constructed inter-document hypertext links in Information Retrieval (IR). The objectives of the experiments described are to obtain evidence concerning the usefulness of querying and browsing automatically constructed IR hypertexts. Links are built by using IR techniques, as these enable rapid, automatic production of hypertexts from a document collection for accessing the collection itself. These tests are carried out in a laboratory environment and through simulation of link browsing. Results of experiments show that browsing has little impact on the retrieval of relevant documents if used in place of querying or relevance feedback methods, though may be practical if used in combination with them.

Keywords: evaluation, statistical methods, hypertext/hypermedia, automatic construction of hypertexts

1. Introduction

1.1. Grouping and linking information retrieval data

Research into grouping and linking of objects dates back to the early days of information retrieval (IR), when it was recognised that the adoption of links between objects helps disclose their content through semantic inter-relationships. Two of the most widespread example techniques for object grouping and linking are relevance feedback (Robertson and Sparck Jones 1976, Rocchio 1971, Salton and Buckley 1990) and clustering, both of documents (Griffitths et al. 1986, van Rijsbergen 1979, Willett 1988) and indexing terms (Sparck Jones 1971, van Rijsbergen 1979, Salton and McGill 1983).

Document clustering techniques provide additional information on document content. The latter may then be described by using the linked documents, rather than with the occurring indexing terms. The same also applies to term clustering. Under the clustering and association hypotheses made for documents and indexing terms, respectively, it is possible to design more effective strategies than that of a linear query-based search. In particular, it is possible to enlarge the set of retrieved documents or expand the query by retrieving clusters instead of individual documents, or by adding clusters of indexing terms to the query, respectively.

Relevance feedback is one of the most renowned search techniques used to effectively improve a classical IR system by providing an additional degree of interaction between users and IR systems. Several relevance feedback algorithms have been proposed (Harman 1992), yet it has been demonstrated that their effectiveness may depend on collection characteristics (Salton and Buckley 1990).

Relevance feedback and clustering are indirect means for linking documents together. The drawback of these techniques is that the degree of interaction is rather low, and that the semantics of these implicit links remains hidden to users, making it impossible to supervise the linking mechanism to adapt it to their own information requirements.

1.2. Hypertexts for information retrieval and automatic hypertext construction

The need to group and link documents and indexing terms has been made explicit and then formalised through the definition of a hypertext IR (HIR) model (Agosti 1988). An IR hypertext system is a network of nodes designed for storing information and establishing interconnections through semantic relationships which are represented as links. Link browsing thus directs users to the desired information being stored in the nodes.

The feature that makes such a hypertext differ from a classical IR system is that the auxiliary data representing the semantic content of the data stored in the nodes are also represented in terms of links. Users explicitly and directly navigate across the links and employ the semantics they represent to express their own information requirements rather than formalising their requests as more or less complex queries. A crucial requirement for HIR systems is that they are expected to help users express their own information needs and direct to the relevant nodes at browsing time. Users may then supervise the linking mechanism by directly selecting links and nodes, without "delegating" the retrieval of relevant data through queries to the system.

The HIR model is considered as a potentially effective alternative to and shares some of the techniques of the classical Boolean, vector-space, or probabilistic model. Indeed, the technique of interconnecting nodes through links and the browsing-based search strategies appear realistically close, for instance, to the classification, indexing, and searching processes carried out by indexers or libraries, as stressed by Sparck Jones and Willett (1997).

Hypertexts for IR may be constructed either manually or automatically. Manual HIR construction is a difficult task due to a number of reasons, and namely:

— the large number of nodes and links makes manual transformation of current text document collections into hypertext unfeasible;
— the individual full-text documents currently available in machine-readable format are often too large to be used as nodes from which extract the data being necessary for the automatic link construction.

To address the issue of the size and number of documents stored in real collections, a number of methods and tools have been proposed for automatic detection of links and extraction of smaller fragments to be used as nodes from documents (Agosti and Allan 1997). Most of the methods proposed for automatic hypertext construction are based on IR techniques since, as stressed above, researchers in the field of IR have always dealt with the automatic representation of semantic relationships between IR objects, i.e., documents and terms (Agosti et al. 1997).

The trade-off between manual and automatic hypertext construction is analogous to that existing between manual and automatic indexing. The main point in favour of automatic HIR

construction is that the reproduction of hypertexts from collections available in electronic form is fast and consistent. However, the greatest drawback of this method is that resulting hypertexts are more primitive than those that are constructed manually. At the current state-of-the-art in automatic construction techniques, hypertexts that effectively represent concepts as nodes and semantic relationships as links are made by the intellectual effort of human experts and, if available, through the support of a semiautomatic tool.

An argument in favour of automatic hypertext construction is that it involves less subjective and partial judgement concerning the quality of hypertext than a human-made construction does, although automatic tools do also incorporate some subjective reasoning, as they are designed and implemented by humans. For example, most automatic hypertext construction methods are based on the computation of the vector cosine to assess the similarity between document nodes. However, this measure is only a rough approximation of the semantic closeness between documents, as has been demonstrated by Allan (1997). Automatic construction helps reduce the risk of inconsistency and missing nodes or links that may occur with a man-made hypertext construction.

1.3. Evaluation of automatically constructed hypertexts for information retrieval

The evaluation of automatically constructed hypertexts is intended as a procedure for assessing the capability of such hypertexts in retrieving information, and is therefore an important step in deciding whether an automatically constructed hypertext is as effective in supporting users retrieving information as with other IR methods (Sparck Jones 1981). As HIR systems are a particular kind of IR system, the procedure of evaluating HIR systems is closely related to the issue of evaluating IR systems. Thus, the evaluation of automatically constructed hypertexts inherits the same weaknesses and strengths of evaluation, though it is a more complex task than for IR systems. A few of the differences are listed below.

1. First of all, HIRs are *networks* of documents and auxiliary data rather than flat collections of plain data. Evaluation must take into account the fact that hypertext links represent semantic relationships between documents and therefore the relevance of a document with respect to a query depends on directly- and indirectly-linked documents.

   Indeed, the dependence between relevance judgements collected during browsing-time is stronger than those during querying-time, as (i) hypertext documents are linked directly, (ii) users do necessarily see documents one after another, and (iii) their judgements about the relevance of a document does strongly depend on the judgements given on the previously seen documents.
2. Evaluation methodologies for IR systems are designed for assessing query-based search processes, whereas HIRs are employed for browsing and querying the document base in an integrated fashion. One should effectively evaluate the new HIR system capabilities in general, and then perform evaluation of the integrated querying and browsing functions that have been automatically developed and made available to the users.
3. Evaluation measures used in IR in a laboratory setting (i.e., precision and recall) are only partially usable when evaluating the characteristics of hypertext information retrieval systems, as these measures are based on the assumption of unlinked documents and

auxiliary data. New measures should be defined to take into account the presence of a network of nodes.

4. When using classical test collections, one has to consider that the judgements collected in a test collection are given with respect to a query representing an information requirement formulated before retrieving documents. On the contrary, users browsing a hypertext formulate their own information requests at navigation-time and therefore judgement regarding the relevance of a visited document is given on the grounds of a partial representation of the request itself. This means that the judgements collected in a test collection for HIR evaluation should represent these relationships.

Evaluation of automatic hypertext construction methods may be approached by measuring three different, complementary factors, as illustrated in a few example research works cited and surveyed in the following:

— the *goodness of nodes in storing relevant data*, as done for example by Salton et al. (1997);
— the *hypertext topology*, as tested by Botafogo et al. (1992), Furner et al. (1996), Blustein et al. (1997), Smeaton and Morrissey (1995);
— the *degree* to which *links are able to represent the semantic relationships between nodes*. As regards this point, the following criteria for assessing the quality of hypertext links have been identified (Thistlewaite 1995):

- *Consistency* between different hypertexts constructed by a specific tool or human experts, as presented in (Blustein et al. 1997, Furner et al. 1996).
- *Soundness* of links in connecting only semantically related nodes. This criterion is analogous to the notion of precision (Croft and Turtle 1993, Savoy 1997, Blustein et al. 1997).
- *Completeness* of links in connecting all relevant nodes as recall is the degree to which all relevant documents are retrieved (Croft and Turtle 1993, Savoy 1997, Blustein et al. 1997, Smeaton 1995).
- *Updating capabilities* of hypertext links by inserting new nodes without violating soundness and completeness (Agosti et al. 1998, Smeaton and Morrissey 1995).

A number of evaluation tests on automatically constructed HIRs have recently been conducted. As in standard IR evaluation procedures, tests on HIRs may be performed within laboratory or operational environments. With respect to the adopted methodology, the research work in HIR evaluation can be broadly classified into two approach methods, namely user and system-sided. The majority of the research contributions concerning evaluation of automatically constructed hypertexts has been approached from the system-side, as:

1. user-sided evaluation is more resource-consuming and may be difficult to repeat, and
2. it is perfectly natural to perform evaluation measures through the same software tool that generated the resulting hypertext.

1.4. User-sided IR hypertext evaluation

The evaluation of HIR effectiveness is defined as *user-sided* when it is carried out through operational or laboratory tests with the active and direct participation of end-users. Some of the proposed evaluation procedures have been designed to compare man-made hypertexts with automatically constructed hypertexts, under the assumption that the former are the "ideal" ones and therefore much "better" than the latter.

Blustein et al. (1997) provide a report on two automatically constructed hypertext evaluation methodologies. One methodology aims to compare an automatic hypertext with an "ideal" one by computing the correlation coefficients between the length of paths between documents and the document similarity expressed through the cosine measure of the weighted vector-space model. The other methodology is based on assessments given by real searchers. It is the latter that belongs to the user-sided category. The experiment is an instance of the scheme described by Tague-Sutcliffe (1992), and it is an important methodological frame of reference for hypertext evaluation. The relevance of that paper is due to the considerations regarding the aspects to be dealt with hypertext evaluation, and in particular:

1. the usefulness of a hypertext may be better understood if it is compared with the corresponding plain texts from which is generated;
2. the population of users for which the hypertext is designed should be known in advance;
3. the significance of tests does strongly depend on the degree of "representativeness" of the sample of users—for example, the more the population is divided, the larger the sample should be;
4. a distinction has to be made between observable variables and external factors—the interface and the main subject of the hypertext may be considered as external factors, as they are not the central focus for hypertext construction and consequently should be controlled;
5. because of inter-document links, judgements regarding the relevance of a document also depend on the assessments concerning previously visited documents.

Another interesting approach to user-sided IR hypertext evaluation is presented by Furner et al. (1996). The authors report on an experimental hypertext comparison study, where several different people were asked to produce a hypertext representation of a full-text document. The aim of the tests carried out by those authors was to measure the inter-linker consistency, i.e., the extent to which different human experts produce different hypertexts. The investigation required the calculation of similarity measures between pairs of manually produced hypertexts. It was shown that the structure of a hypertext document is crucially dependent upon the person who created the links. This important experimental result implicitly supports the study and development of methods for automatic authoring of hypertexts, as the application of a sound automatic authoring method can produce a hypertext that does not incorporate just one specific subjective design and development view, and is reproducible starting from the same initial collection of flat documents. The authors of the study conclude that they were "unable to reject the null

hypothesis that there is no positive association between inter-linker consistency and retrieval effectiveness".

Most of the research work on automatic hypertext construction is focused on evaluating the quality of links. An important methodological contribution for evaluating the quality of nodes is presented by Salton et al. (1997), in which methods for automatically extracting summaries from text documents are described. The extracted summaries may generally be used as hypertext nodes. The authors set a test procedure to compare manually extracted summaries to automatically extracted ones. Users submit a document to the system that presents two summaries, one manual and one automatic one. The user judges both summaries with respect to his or her own notion of quality. Under the assumption that manually made summaries are the "ideal" ones, the overlapping of selected automatic summaries with manual ones is computed as a measure of goodness of the automatic summary.

1.5. System-side IR hypertext evaluation

The evaluation of HIR effectiveness from the *system-side* is based on the computation of quantitative measures and/or the simulation of retrieval processes, such as querying and browsing. These tasks are performed through the use of a number of test collections within a laboratory environment and without involvement of any real users. This kind of evaluation may include:

— the computation of classical precision-recall measures (Croft and Turtle 1993, Savoy 1997);
— the computation of measures describing the hypertext topology (Botafogo et al. 1992, Furner et al. 1996, Blustein et al. 1997).

Croft and Turtle (1993) performed a comparison between a heuristic spreading activation strategy and a probabilistic retrieval model incorporating inter-document links. The main findings are that the use of hypertext links makes retrieval more effective than strategies without links. Specifically, manually constructed links, such as bibliographic citations, are more effective than automatically constructed ones, such as the nearest neighbour links. The authors stress the importance of implementation issues, as the use of hypertext links in retrieval strategies requires additional computation resources to store and process links.

Savoy evaluates the effectiveness of inter-document links designed and implemented on the grounds of three kinds of bibliographic citations, i.e., bibliographic references, bibliographic coupling, and co-citations, as well as links based on nearest neighbour nodes (Savoy 1997). Two test collections, i.e., CACM and CISI were employed (Fox 1983). The former also included bibliographic citations between documents. The results confirmed the findings made by Croft and Turtle (1993) and demonstrated that links based on bibliographic citations are more effective than links based on nearest neighbours, as the former are carefully inserted by the document authors. The effectiveness of bibliographic data depends on the availability and organisation of references and citations. For example, Crestani and Melucci (1998) conducted a case study on automatic authoring by transforming a textbook into a hyper-textbook. A textbook, such as that used in (Crestani and Melucci 1998), is often

full of bibliographic data though they consist of out-going citations, and few bibliographic references are co-cited by the same text paragraphs. The same applies to journalistic articles.

The important lesson to be learnt from the work presented in (Savoy 1997) and (Croft and Turtle 1993) is that man-made links (such as bibliography-based links), if available, are more effective than, and are an "upper limit" to automatically made links. Therefore, it is necessary to dedicate more research work on the evaluation of automatically constructed links to understand whether they are effective enough whenever links made by a human expert are absent, or in what proportion the effectiveness of automatic links is less than that of manual links.

Another approach for evaluating the quality of a hypertext is based on the computation of measures describing the hypertext topology. An example of such an approach is described by Botafogo et al. (1992) to evaluate the "goodness" of a hypertext that is being manually constructed by a human expert and based on hierarchical structures. The interest in hierarchical hypertext structures is based on the fact that they are considered as the most effective means for avoiding disorientation due to too many jumps when navigating the hypertext. The evaluation methodology is based on two items of information: (1) the indications given to the author to recover lost hierarchies and define new ones, and (2) functions measuring the compactness and stratum that are respectively the "intrinsic connectedness of the hypertext", and "the degree with which the hypertext is organised so that certain nodes must be read before others". The incorporation of topology-based measures has also been proposed by Furner et al. (1996). However, their aims are different from those reported in (Botafogo et al. 1992), as they employ several graph theory results, such as adjacency, distance and converted distance to compare the hypertexts produced by different authors.

Though the methodologies proposed in (Botafogo et al. 1992) and (Furner et al. 1996) are designed for manually constructed hypertexts, they may be incorporated in any automatic construction algorithm. Indeed, a topology-based routine may drive an automatic hypertext construction prototype towards a "good" hypertext, where the degree of "goodness" is expressed, for instance as measures being computed on adjacency matrices.

As mentioned above, Blustein et al. (1997) presents an evaluation method based on the computation of numerical measures. In particular, the authors studied the correlation between the similarity function used to decide whether to insert a link between two documents and the length of paths between documents, and formulated the following conclusions: (1) similarity value and path length are inversely correlated, i.e., highly similar documents are connected by short paths, and (2) correlation decreases as the number of out-going links increases. This means that considering document nodes with a single out-degree leads to sound, but incomplete hypertexts. On the contrary, nodes with several out-going links generate complete, though imprecise hypertexts.

1.6. Methodologies and tools for automatic construction of hypertexts for information retrieval

A methodology for automatic hypertext construction was proposed in (Agosti and Crestani 1993) and has been further implemented in a prototype system called Tachir (a Tool for Automatic Construction of Hypertexts for Information Retrieval) (Agosti et al. 1996).

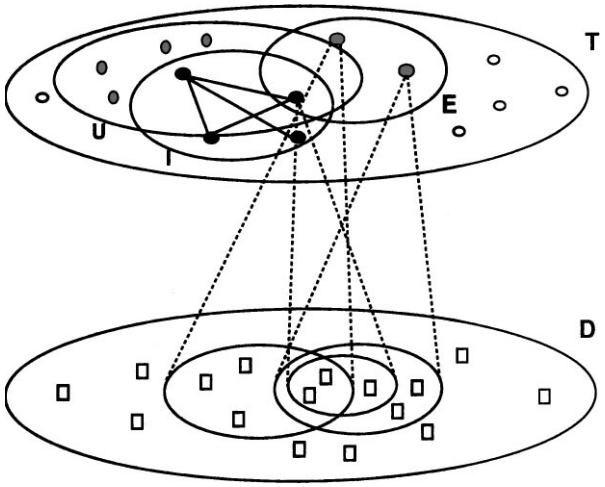

Figure 1. The Explicit model.

Tachir employs IR techniques, such as term probabilistic weighting and similarity functions, to automatically detect document-to-document and term-to-term links, as well as links between documents and terms.

Tachir is based on a two-level conceptual HIR model called Explicit (Agosti et al. 1991), depicted in figure 1, and represented in the three-level schema of figure 2, which stresses the different types of nodes and links involved in the automatic construction process and implemented by the architecture depicted in figure 3, where links between nodes are implemented as ranked list of anchors. Ranking values are computed by using similarity functions or index term weights. For example, a link between two documents is detected through a similarity function, which determines the insertion of a link if the similarity value exceeds a stated threshold. As relationships between documents are of the "many-to-many" kind, links are implemented as lists of anchors, where an anchor corresponds to a similar document and the list is associated to the starting document. Anchors are ranked by their similarity function value. Similarly, links between terms and between documents and terms are implemented by using term similarities and weights, respectively.

Tachir works as the unique author on the grounds of a conceptual schema given by the Explicit model and automatically builds links accordingly to the implementation architecture of figure 3. In this way, hypertext *consistency* is guaranteed. The ranking list of anchors that implement links between nodes helps support *soundness* and *completeness*, as similarity functions or index term weights enable threshold setting. The anchor lists also serve as means for suggesting which are the most relevant document nodes or the most useful term or concept nodes of the Explicit model. One of the greatest limitations of Tachir was the difficulty in updating, as all newly inserted documents required a complete reconstruction of the hypertext. A system called Ofahir (On-the-Fly Automatic authoring of Hypertexts for IR), presented by Agosti et al. (1998), solved the problem of hypertext *updating* as the basic data used to automatically build hypertexts are stored in an object-oriented database. Any addition of documents only requires the construction of links starting from and ending

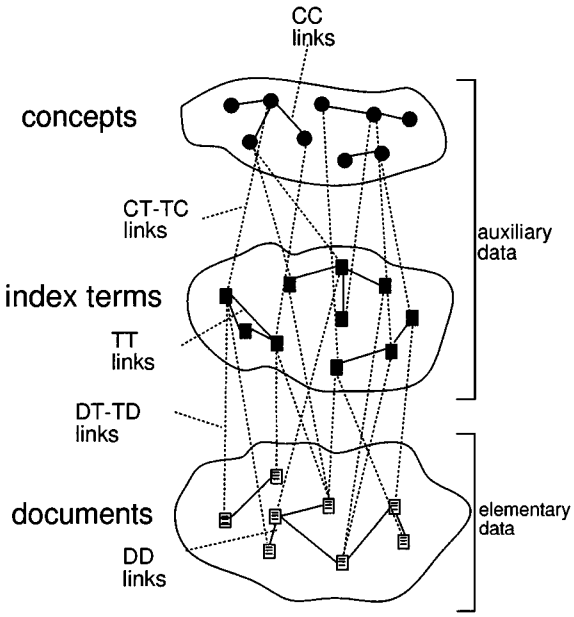

Figure 2. The conceptual Tachir schema.

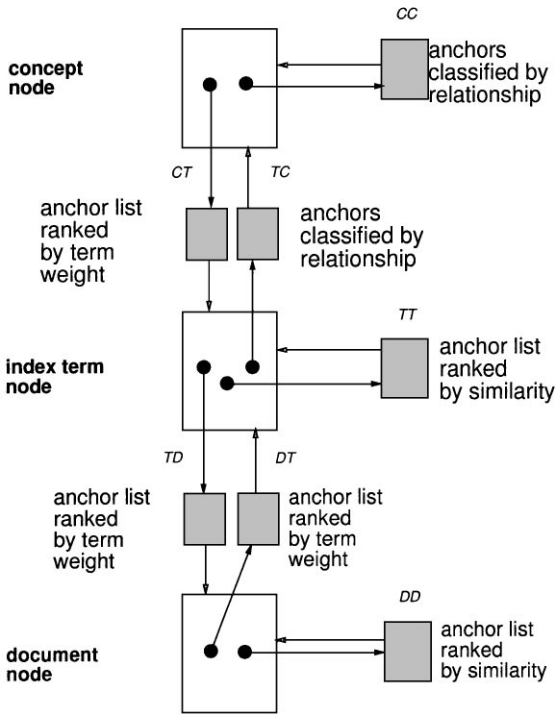

Figure 3. Hypertext implementation.

at the new document, as well as the construction of links involving the indexing terms of the new document. Querying and browsing are implemented on top of the same database. This means that users may query the document collection and start browsing from the retrieved documents by following the links between similar documents and related indexing terms.

1.7. Querying and browsing a hypertext for information retrieval

The main functions provided by a system implementing a hypertext for IR automatically constructed by Tachir include browsing and, if the more recent Ofahir is used, querying as well. Through the querying function, the system retrieves a set of document nodes storing data with a semantic content that is relevant to a user's information requests. The document nodes are organised as a list ranked by a measure of relevance and presented to the end-user, who may start browsing using the documents as entry points. The current implementation employs a probabilistic IR model to rank retrieved documents. If a user was using a classical IR system and the retrieval result was unsatisfactory, the system should offer two possible alternatives, namely to modify the query manually by adding or removing terms, or to perform a relevance feedback operation, if available.

Figure 4 represents the document space using the graphs proposed in (Salton and McGill 1983) to give an idea of documents, queries and their relationships. The measures of relevance of documents with respect to a query are represented as arcs. The arc length is inversely proportional to a retrieval status value. The query is the centre of a radius including documents having a retrieval status value that exceeds a given threshold, and defined as a "retrieval radius". If the maximum number of nodes to be included within a radius is fixed, the retrieval radius corresponds to the lowest retrieval status value. The graph on the left (A) depicts a query and its corresponding retrieval radius. The one on the right (B) illustrates

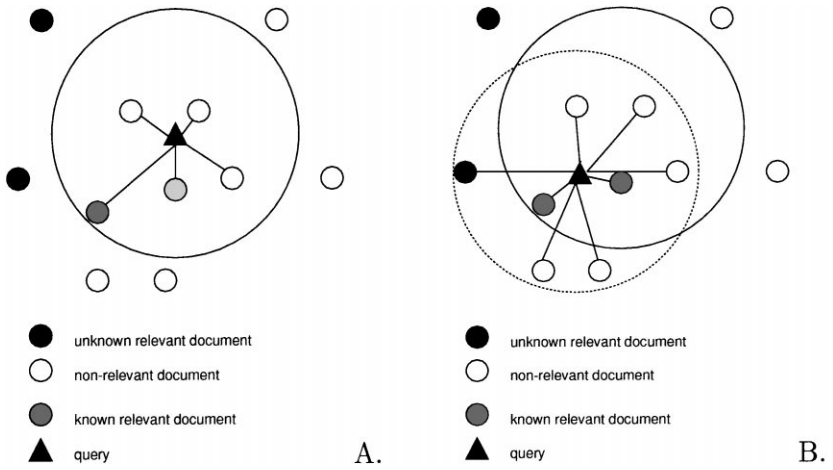

Figure 4. The space of documents and queries.

the query being modified after relevance assessments have been used to move it towards relevant documents.

In general, unsatisfactory retrieval results are more a rule than an exception, especially if no relevant documents are retrieved within the list that is presented by the system to the end-user. A particular case of ineffective retrieval is one that occurs whenever short queries are used, where the notion of "query length" may be empirically represented by the number of query terms. The use of search engines running under Internet has given highlighted the problem of short queries, i.e., queries being expressed with a number of terms lower than the usual average number of query terms of test collections. In general, the user of a search engine typically types a few words and asks the system to interpret their meaning, often avoiding the mental effort of choosing the best terms to express his or her information requirement. Most of the statistical retrieval techniques developed and evaluated so far are based on long queries. Kwok (1996) stresses that the function provided by long queries is twofold, namely they limit the search environment by reducing the degree of ambiguity, and they provide an indication of how much a query term is important in describing a topic of interest. The main drawback of short queries is that they make discrimination between relevant and non-relevant documents more difficult than with longer queries, as they are unable to limit the search space, and they provide poor statistical indications regarding the importance of terms. Kwok (1996) proposes an automatic method for estimating the importance of query terms as a means for making short queries more document-discriminating. An alternative method would be based on browsing the space of terms, if inter-term relationships were provided.

If the system is implemented on top of a HIR, users may take advantage of the browsing-based functions offered. A hypertext like the one automatically constructed by Ofahir is capable of supporting the user during retrieval whenever a query-based retrieval gives unsatisfactory results. This may be done through browsing, which is the second function provided by the hypertext system we are considering. The user marks a retrieved document as relevant, and from that document he or she browses other "similar" ones. The purpose of similar document browsing operations is to retrieve relevant and previously non-retrieved documents by navigating the links that represent the relationship associations between documents. The links being detected on the basis of document similarity functions and browsed by the user may help increase retrieval effectiveness.

Figure 5 shows the difference between document browsing (B) and retrieval (A). Similarity links between documents are represented as arcs between circles. The arc length is inversely proportional to the strength of inter-document similarity. If we consider a document from which browsing can take place, the set of all the equally similar documents determines a circle of which the document in question is the centre. The dashed radius represents the set of documents that are similar to the centre of the radius itself, this is defined as the "similarity radius". The smallest retrieval radius includes documents, say first k out of the total n retrieved having a higher retrieval status value than the threshold represented by the radius. The largest retrieval radius includes the $n - k$ documents with a lower retrieval status value than the first k.

The effectiveness of document similarity link browsing may depend on the degree with which the cluster hypothesis or clustering tendency characterises the actual document

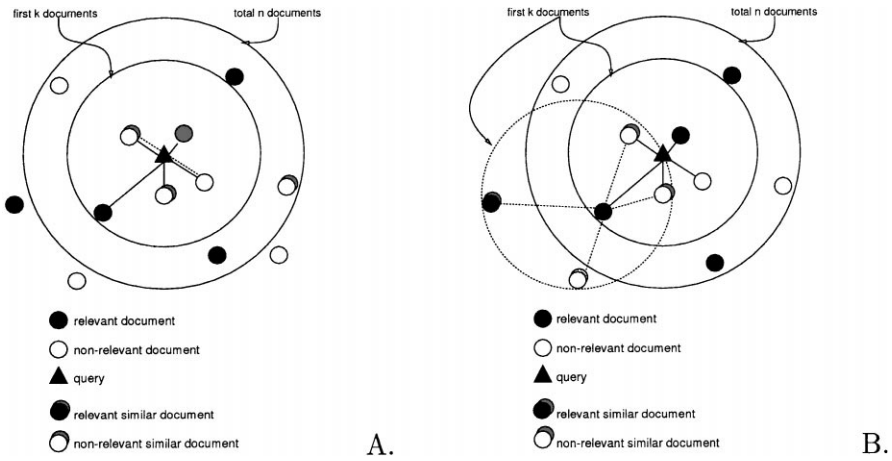

Figure 5. Document browsing vs. retrieval: Starting from centre and starting from boundary.

collection from which the hypertext has been built. If the relevant documents tend to be uniformly "dispersed" throughout the document space, i.e., if the cluster hypothesis is rejected, the effectiveness is likely to be poor, as few relevant documents will be linked through similarity links to the one from which navigation has started. On the contrary, browsing may be effective if relevant documents tend to be clustered together and separated from non-relevant documents. Previous research work (Griffitths et al. 1986, Jardine and van Rijsbergen 1971, van Rijsbergen and Croft 1975) regarding document clustering extensively investigated such a relationship, even though within a classical IR environment without HIR browsing facilities. The relevance of this latter work to the previous work on document clustering and cluster hypothesis is due to the common inter-document similarity-based linking device that is at the basis of Tachir or Ofahir hypertexts.

The effectiveness of link ranking may be related to the degree with which high retrieval status values correspond to high proportions of newly retrieved relevant documents. In other words, ranking both retrieved and similar documents by using the retrieval status value and similarity, respectively, is likely to have a direct impact on browsing effectiveness. The underlying assumption is that the more a document is judged by the system as relevant to the query, the higher the chance of finding other relevant documents that are similar to the selected one. Therefore, the more a user selects the top-ranked retrieved documents from which a browsing process starts, the higher the browsing effectiveness is expected to be.

2. Experiments

In this paper, the author presents the design and results of several experiments carried out to evaluate a hypertext automatically built by Tachir or Ofahir according to the document level of the conceptual architecture illustrated in figure 2. The main objective of the experiments was to evaluate the effectiveness of integrating browsing and querying functions into such an IR hypertext. The test hypotheses are listed below:

— To evaluate whether inter-document similarity link *browsing* would lead to the retrieval of more new, relevant documents than those that would be retrieved if a user chooses to perform a query-based search;
— To evaluate if *ranking* by a document similarity function is effective, i.e., if top-ranked links are entry points for retrieving (by browsing) a higher number of relevant documents than the bottom-ranked links.
— As described in Section 1, the degree of *clustering tendency* may influence the effectiveness of document similarity-based techniques. This suggests the necessity to evaluate whether such tendency influences the effectiveness of searches based on document similarity link browsing.
— An end-user may access an IR hypertext through querying, browsing, or both. Combining querying and browsing techniques means that the end-user performs both query-based and browsing-based retrievals. Therefore, another hypothesis to be tested is whether the *combination* of querying and browsing functions helps improve the overall retrieval performance with respect to querying alone.
— In Section 1, it was stressed that *short queries* cause a drop in retrieval performance. This would require evaluation of whether short queries determine lower effectiveness of both querying and browsing retrieval, and if browsing can help improve querying results by providing effective entry points to non-retrieved material.

2.1. Experimental methodology

The mechanisms underlying the process of browsing hypertext links are fairly complex to describe. This complexity is due to the intrinsic interactive nature of hypertexts, that is to the presence of a human user that makes the modelling of a browsing process very difficult to formalise and simulate. One needs to control the variables of interest and isolate the external factors, such as the user-hypertext interface. This requirement can hardly be met if tests are performed with the users' involvement, for which one should implement a working prototype to perform an evaluation with real searchers. In particular, an appropriate user interface would need to be designed and implemented which could make the interface itself a determining factor in the final results. Blustein et al. (1997) also point out that "it is not always practical to evaluate hypertexts with real readers because the documents are often too complex for people to explore fully and objectively".

The experiments were carried out in a laboratory environment and without the participation of real searchers. In order to make results reproducible, several programs simulated simple browsing paths and computed a number of performance measures with different test collections. This experimental design choice was determined by the need to test the basic algorithms for the automatic hypertext construction algorithms at the current stage of development, as they constitute the engine of Tachir and Ofahir. The results regarding the effectiveness would provide useful data for modifying the hypertext architecture and the automatic construction methodology accordingly.

The evaluation presented in this work attempts to take into account the fact that a hypertext is mainly used for browsing, although the tests are performed within a laboratory environment and without real searchers. Due to the complexity of modelling all the possible

browsing processes, it was necessary to identify a short browsing path that could be performed by a generic user to retrieve information. The identified path was the basic element of more complex paths consisting of the combination of many other short ones. Therefore, the effectiveness of short browsing paths were evaluated to extend the results to more complex browsing processes. Moreover, the experiment concentrated on the first level of the architecture, i.e., the document level, to isolate the variables influencing the effectiveness of browsing through document similarity-based links and control the complexity of studying the factors related to the other hypertext levels together.

2.2. Search tactics and tests

The experiments involved the comparison of three *search tactics* the user exploits when using a hypertext:

1. *Linear* search tactic: the user submits a query and browses all the n retrieved documents presented within the initially ranked list. The following tests were performed to evaluate the linear search tactic:
   - — a baseline test without relevance data,
   - — a relevance feedback test with relevance data collected from the first $k = n/2$ documents retrieved through the baseline search. For each query, index weights were modified by considering the relevant documents within the first k ones.
2. *Hypertext* search tactic: the user performs a linear search without activating relevance feedback, and stops browsing after having seen k documents; a document is then selected and a new browsing process is started to retrieve k additional similar documents.
3. *Combined* search tactic: the user performs a linear search tactic and examines all the n retrieved documents. The user then returns back to the first k documents and starts a hypertext search tactic to retrieve n similar documents.

The purpose of comparing these search tactics is to evaluate the search alternatives a user may choose within a hypertext IR environment. Indeed, the user may sequentially browse the entire list of retrieved documents, as the linear search tactic would simulate, or alternatively navigate the hypertext using links, as the hypertext search tactic would simulate. The combined search tactic would simulate a user performing a search procedure consisting of both tactics.

With a linear search tactic, the user would be able to retrieve as many relevant documents as there are in the n-documents list. With the second tactic, the user retrieves the relevant documents found in the first half of the k-documents list and then the relevant documents found during the subsequent browsing processes. The first half of the document list reached through browsing is scanned in place of the second half of the initial document list. The total number of examined documents is then the same, but in the second tactic the user follows the hypertext links instead of examining the whole initial document list. With the combined search tactic, the user exploits both methods to retrieve relevant documents by examining $n + n$ documents. In terms of the number of documents examined, the combined

search tactic would be more time-consuming than performing either a linear or hypertext search tactic, though more effective.

This experimental study differs from past investigations on Cluster hypothesis and on cluster-based searching, because the non-query-based search tactics being tested, whether hypertext-based or combined, have been designed to evaluate browsing. The experiments regarding the combined search tactic would actually give some insights about the contribution of browsing to querying results.

2.3. The selection of entry points for browsing

The experiments intended to simulate a user scanning the list of retrieved documents and marking a number of documents considered as candidate entry points for browsing. The collection of relevance assessments is illustrated in figure 6 and described below.

1. The first half list (A) of k retrieved documents is divided into h groups of k/h documents each. A is scanned and one document is selected per each group by marking h documents as entry points in total. Either the top-ranked relevant document, if any, or the top-ranked document of each group is selected as the entry point.

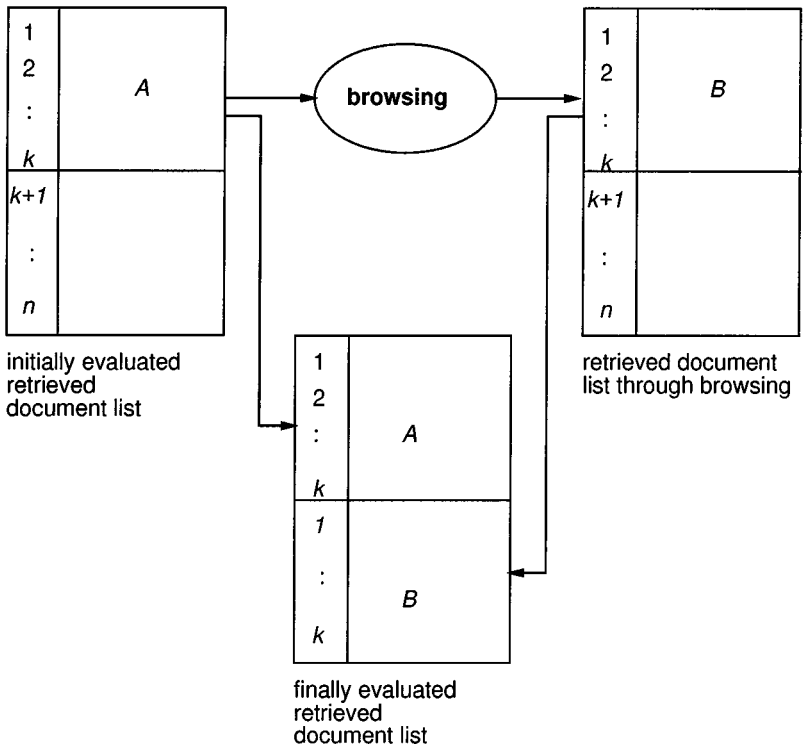

Figure 6. Merging the browsing-path-based retrieval results to the query-based retrieval results.

2. A similarity-based link is created from each selected document to connect a list of n similar documents. Each list of similar documents is ranked by the corresponding similarity measure. These lists are not merged, but are alternative to each other, and therefore h alternative lists of similar documents are created during the experiments for each marked document respectively.
3. With the hypertext search disabled, the k most similar documents (B) replace the k bottom-ranked documents retrieved beforehand. Therefore, a new ranked retrieved document list is created as a result of the concatenation of A and B. The new concatenated and the initial lists are then compared to check whether and to what extent the quantity of relevant similar documents differ from each other. With the combined search tactic, B does not replace, but rather is examined after the retrieved documents in A.

In the experiments reported in this paper, n, and h are set to 40 and 4, respectively. Each group consists of 5 documents. The author chose low values for n and h, as the aim of this experiment was to simulate a hypertext IR system that presents users with a list of items that is short enough to be displayed on a screen. Greater values would be unrealistic, as more than a few dozens of links cannot fit on a screen, or may be of little use to the searcher. Setting these low values would stress the importance of retrieving as many relevant documents as possible from several small fractions of documents linked through similarity-based and automatically constructed links.

2.4. Test collections

As no ad-hoc test collections for HIR evaluation were available, classical test collections had to be used. Six test collections were employed to carry out the experiments and are described in Table 1.[1] The size of the chosen collections is of different orders of magnitude. In particular, the 1987 Ohsumed collection, which is a part of the complete Ohsumed set (Hersh et al. 1994) is the largest used data set and the Medline is the smallest one.

Each collection was automatically indexed by describing each document as a binary vector of index terms. Each index term i was weighted according to the probabilistic

Table 1. Collection details.

	Collection					
Parameter	Ohsu	NPL	CACM	CISI	CRAN	Med
Documents	54710	11429	3204	1460	1400	1033
Queries	106	93	64	112	225	30
Relevant documents per query	7.45	22.43	12.48	27.83	8.18	23.30
Queries with no relevant documents	9	0	12	36	0	0
Index terms per document	53.13	18.58	21.83	47.61	52.60	51.84
Index terms per query	6.84	6.92	9.23	28.80	9.07	10.63
Index term density test	.0012	.0025	.0043	.0036	.0131	.0061

weighting scheme $f_{ij}(\log p_i/(1-p_i) - \log q_i/(1-q_i))$ where f_{ij} is the frequency of term i in document j, $p_i = (r_i + 0.5)/(R+1)$, $q_i = (n_i - r_i + 0.5)/(N - R + 1)$, n_i is the number of documents indexed by term i out of the total N documents, and r_i is the frequency of term i within the R relevant documents that had been marked during a relevance feedback process. If no relevant documents were marked, $R = r_i = 0$ for each term i (van Rijsbergen 1979, Croft and Harper 1979, Robertson and Sparck Jones 1976). Queries were automatically indexed as well, and binary weights were assigned to query terms.

2.5. Short queries

In Section 1, it was pointed out that the most common form of interaction between a user and an IR system (such as a Web search engine), is performed by using rather short queries. One question this paper intends to investigate is whether the use of a hypertext search tactic is an effective means for improving retrieval effectiveness in this particular kind of situation. To test this hypothesis, the author performed several tests by reducing queries of the employed test collections before executing the tests again. The reduced set of query index terms would describe a query from which some of the index terms were excluded by the user from his or her own first formulation. The reduced query would then represent the same information request, but in a more imprecise and incomplete manner. Query reduction was performed on the set of index terms describing the query, and therefore any original non-index terms, such as stop-words, were excluded from the reduction process. Query reduction was made so that a few keywords would always be present. The reduced set of query keywords would thus correspond to the indexed version of a real short query. In order to test the impact of different average query lengths on effectiveness, the author considered three percentages of reduction: 10, 50 and 80%. Query reduction was performed by "randomly" removing the given percentage of index terms. The term "randomly" refers to the fact that index terms were randomly selected and removed from the set of terms describing the query.

2.6. Measures

Retrieval effectiveness was measured through the classical recall and precision ratios computed for eight groups of five retrieved documents, i.e., after retrieval of groups of 5, 10, 15, 20, 25, 30, 35, and 40 documents. Recall and precision ratios were computed for each query and averaged over the total number of queries. As the hypertext search tactic is alternative to a linear search, the recall and precision values computed at 5, 10, 15, and 20 documents retrieved through browsing were indirectly compared to the values computed at 25, 30, 35, and 40 documents retrieved through the initial query. More specifically, recall and precision were computed for the $A + B$ concatenated list and compared to the ones computed for the whole retrieved document list of which A was the first half. These computations were made to compare the effectiveness of the sequential scanning of all retrieved documents to the effectiveness of the hypertext navigation of links between the top-ranked retrieved documents and the most similar browsed documents. The F-measure (van Rijsbergen 1979) was computed to summarise the retrieval performance measured by

recall and precision. The F-measure was expressed as $F = 1/(\alpha P^{-1} + (1-\alpha) R^{-1})$ where P stands for precision, and R stands for recall. The weight given to precision and to recall in contributing to F-measure is represented by α. The higher the α, the more the precision is important in the F-measure. For the combined search tactic, the quantity of relevant documents retrieved from the concatenated document list was computed as a percentage, but not from the entire previously retrieved document list of which A is the first half.

In query-based search strategies, a user would retrieve as many relevant documents as possible by limiting the proportion of non-relevant documents. Therefore, the ideal result would be achieved if recall and precision are maximised. Through relevance feedback or query expansion, if earlier retrieval results are unsatisfactory (Efthimiadis 1996), yet the degree of interaction is lower than offered by HIR, then it would be possible to increase recall and precision. In browsing-based search strategies, users navigate the document environment to retrieve relevant documents by following the inter-connecting links. The proportion of relevant documents that are retrieved changes as navigation proceeds. Once the user reaches a list of document links, he or she can judge them for their relevance, and if no relevant documents are found, he or she can carry on searching by selecting several documents as entry points to browse the list of links to other similar documents. It is therefore important that the retrieved list includes as many links to relevant documents as possible, as a relevant document is likely to direct the user further to additional relevant ones. Precision should therefore be as high as possible at the level of individual navigation steps. On the other hand, recall can be increased by navigating the document space, as links may help the user to retrieve additional relevant documents. It is therefore assumed that the precision of hypertext search tactics at individual navigation step level is more important than recall. In the experiments, the author consequently set $\alpha = 0.67$. Therefore precision was assigned a weight double that of recall.

Indexing term density measure was computed for each collection to estimate its clustering tendency. Indexing term density measures the degree of sparseness of the co-occurrence matrix without using relevance information, as described by Willett (1988), together with other clustering tendency measures. This test is given by the total number of postings divided by the product of the number of documents and the number of indexing terms. This measure was adopted to estimate the clustering tendency, as it does not require the knowledge of relevance data, as the overlap test does (van Rijsbergen 1979); nor does it require the computation of similarity values, as the nearest neighbour test does (Voorhees 1985). As suggested by Willett (1988), "term density was found to correlate best with the effectiveness of cluster searching". As cluster searching is based on inter-document similarities, as in the non-linear search tactics, the Author preferred to use the indexing term density test. High density test values would indicate that inter-document similarity-based searches are likely to produce useful results.

2.7. Results

Experimental results regarding the three search tactics for all the test collections and complete queries are reported in Table 2. The Table is divided into three parts to present the results for each search strategy. The cells of the first and second parts report F-measures.

Table 2. Experiment results with complete queries.

	Collection					
Search tactic	Ohsu	NPL	CACM	CISI	CRAN	Med
Linear search						
Baseline	.155	.208	.173	.159	.269	.483
RF	.167	.211	.203	.137	.276	.542
Hypertext search						
after 5 docs.	.159	.204	.178	.158	.273	.504
after 10 docs.	.155	.205	.172	.155	.266	.495
after 15 docs.	.150	.203	.172	.155	.266	.490
after 20 docs.	.150	.203	.171	.153	.266	.483
Combination						
after 5 docs.	10.1%	5.5%	7.6%	4.8%	8.9%	14.0%
after 10 docs.	5.2%	6.1%	6.8%	4.9%	6.3%	11.2%
after 15 docs.	3.1%	5.9%	6.9%	4.8%	5.9%	12.6%
after 20 docs.	3.2%	5.3%	7.3%	3.2%	5.9%	11.2%

The reported F-measures are the average of the F-measures computed after 5, 10, 15, 20, 25, 30, 35 and 40 retrieved documents. The first part concerns the linear search tactic and reports the results of the three tests: the baseline and the relevance feedback test. The second part represents the hypertext search tactic and reports the results classified for four different browsing entry points. The phrase "after m docs." means that the entry point was chosen between the $(m - k/h)$-th and the mth retrieved document. The third part refers to the combination of linear and hypertext search tactics and reports the percentage of relevant documents retrieved within the similar documents using the hypertext search tactic, though they were non-retrieved among the documents using the linear search tactic. That percentage measures the increase in the quantity of relevant material that a user is able to retrieve if a combined search tactic is chosen rather than a linear search tactic alone.

With respect to the list of hypotheses listed at the beginning of Section 2, the following conclusions may be drawn:

— The results of linear and hypertext search tactics show that, in the case of complete queries, document similarity-based link *browsing* leads to the retrieval of less relevant documents than those retrieved if users choose to perform a linear-based search. In particular, a linear search based on relevance feedback, which is performed on the first 20 documents, is more effective than the hypertext search tactic, except for the CISI collection.
— The results of hypertext search tactics show that link *ranking* has little impact on retrieval effectiveness. The reasons for this may be the low number of retrieved documents and the small size of the retrieved document list from which the browsing entry points are

selected. It is likely that the top-ranked documents, whether relevant or non-relevant, are in most cases equivalent entry points for hypertext searching.

— Test results referring to non-linear search tactics, whether hypertext or combined, would indicate that the *clustering tendency* only partially influences retrieval effectiveness. If the differences between the baseline result and the hypertext search result are computed as percentages, it may be possible to observe that the relationship between the term density test and the usefulness of browsing inter-document similarity-based links is confirmed for the Med collection. However, the 1987 Ohsumed collection shows a positive difference, corresponding to a low term density.

— The results of the search tactic based on the *combination* of linear and hypertext search tactics show that significant improvement is obtained. For example, the quantity of relevant documents retrieved from the 1987 Ohsumed collection by following a link starting from the first 5 retrieved documents amounts to 10.6% of the total set of relevant documents. The latter results indicate that similarity-based link browsing leads to a different set of relevant documents than by querying, even though that set is smaller than the one retrieved through a linear search. The results are consistent with the findings reported in (Griffitths et al. 1986) regarding document clustering, where the combination of the nearest neighbour-based clusters and linear search techniques is suggested.

  The two highest percentages of new relevant documents retrieved through the combined search tactic correspond to the two highest term density values. This result would indicate that the possibility of linking documents according to their similarity yields useful results. However, the Ohsumed collection shows an opposite result, though only partially, as the low term density value corresponds to an increase of effectiveness in using top-ranked documents as browsing entry points.

— Tables 3–5 report the experimental results carried out after the *query reduction* by 10, 50 and 80%, respectively. The order of magnitude of the reported values decreases as the percentage of reduction increases. The results regarding the hypertext and the combined search tactic are directly related to the results of the linear search tactic. This means that as linear search effectiveness decreases, so does the performance of hypertext search operations through document links. This relationship suggests that the additional effectiveness provided by a hypertext search tactic, such as the one described in this paper, depends on the effectiveness of the linear search result from which browsing starts.

  The difference between hypertext search tactic effectiveness and linear search tactic effectiveness increases with greater query reduction factors. This means that hypertext search tactics may be of help in reducing the loss of effectiveness in "short query" situations. The fact that there is still an increase in the number of additional relevant documents after having performed a combined search initiated from the "poor" result provided by a short query indicates that browsing may be able to compensate for the loss of effectiveness. Therefore, the combination of querying and browsing tactics can be justified and suggested. However, it must be noted that the very low level of performance obtained by using very short queries may be improved by adopting other alternative term weighting schemes, such as those reported in (Kwok 1996).

Table 3. Experiment results with queries being shortened by 10%.

	Collection					
Search tactic	Ohsu	NPL	CACM	CISI	CRAN	Med
Linear search						
Baseline	.147	.205	.171	.148	.264	.487
RF	.161	.213	.203	.126	.272	.571
Hypertext search						
after 5 docs.	.151	.202	.178	.148	.268	.509
after 10 docs.	.147	.204	.169	.147	.262	.510
after 15 docs.	.142	.199	.170	.144	.263	.492
after 20 docs.	.142	.199	.173	.145	.261	.491
Combination						
after 5 docs.	10.6%	5.6%	8.9%	4.8%	9.3%	14.6%
after 10 docs.	5.0%	6.9%	6.1%	5.1%	6.8%	13.6%
after 15 docs.	4.0%	6.7%	7.9%	4.2%	6.6%	12.9%
after 20 docs.	3.5%	4.9%	6.8%	3.5%	6.8%	11.6%

Table 4. Experiment results with queries being shortened by 50%.

	Collection					
Search tactic	Ohsu	NPL	CACM	CISI	CRAN	Med
Linear search						
Baseline	.053	.110	.136	.098	.188	.377
RF	.058	.111	.157	.079	.188	.455
Hypertext search						
after 5 docs.	.055	.114	.138	.101	.196	.403
after 10 docs.	.056	.111	.135	.100	.190	.397
after 15 docs.	.055	.108	.131	.097	.189	.405
after 20 docs.	.054	.107	.132	.098	.185	.395
Combination						
after 5 docs.	5.1%	5.0%	4.2%	3.2%	13.4%	17.0%
after 10 docs.	6.1%	7.9%	4.0%	3.7%	10.1%	16.1%
after 15 docs.	3.7%	5.3%	5.0%	3.7%	9.2%	11.6%
after 20 docs.	3.2%	4.4%	7.1%	4.3%	t6.4%	13.6%

Table 5. Experiment results with queries being shortened by 80%.

	Collection					
Search tactic	Ohsu	NPL	CACM	CISI	CRAN	Med
Linear search						
Baseline	.024	.054	.060	.040	.118	.210
RF	.006	.053	.060	.034	.118	.238
Hypertext search						
after 5 docs.	.027	.057	.057	.041	.128	.241
after 10 docs.	.024	.058	.060	.041	.121	.235
after 15 docs.	.020	.055	.062	.044	.120	.235
after 20 docs.	.021	.054	.058	.041	.116	.234
Combination						
after 5 docs.	4.1%	5.7%	3.9%	3.8%	15.1%	16.7%
after 10 docs.	4.7%	6.9%	2.7%	3.7%	9.4%	16.6%
after 15 docs.	4.9%	4.9%	5.4%	3.1%	8.7%	10.7%
after 20 docs.	2.9%	3.8%	6.0%	4.0%	8.0%	15.7%

3. Conclusions

The use of IR techniques (and specifically of statistical and probabilistic techniques) to automatically construct links between documents, such as those underlying Tachir, may prove to be in principle an effective and necessary means for automatic construction of hypertexts, as they enable the fast production of these networks to be used as support for retrieval through browsing. Their properties are indeed known well enough to understand when and whether they work. However, the experimental results discussed in this paper would indicate that IR techniques are inadequate for building effective hypertexts that are supposed to retrieve documents using the links existing between them.

The use of inter-document similarity-based links to search relevant information by using automatically constructed hypertexts has been shown to provide little difference from query-based searches. However, browsing these links helps to retrieve additional relevant documents previously not retrieved by a query-based search. This would suggest that a combination of search tactics is advisable.

It is likely that hypertext browsing, together with querying and relevance feedback, may lead to a significant improvement in retrieval effectiveness. This integration of techniques becomes more and more essential as the performance of query-based retrieval decreases. Further research efforts are therefore needed in this direction. A multi-level hypertext, such as the one illustrated in figure 2, may be more effective thanks to it greater richness of link and node types.

The experiments presented in this paper are to be intended as complementary, rather than alternative to those we could carry out with real searchers. The design and implementation

of tests based on the employment of real searchers would be necessary to fully evaluate an interactive system such as the one based on an automatically constructed hypertext.

Acknowledgments

The author thanks the anonymous reviewers and Professor Maristella Agosti for helpful comments and suggestions on this work. This work was partially supported by the INTERDATA project from the Italian Ministry of University and Scientific Research and University of Padova.

Note

1. "Ohsu" stands for the 1987 Ohsumed collection.

References

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

Agosti M and Allan J, eds. (1997) Special issue on methods and tools for the automatic construction of hypertexts. Information Processing & Management, 33(2).

Agosti M, Benfante L and Melucci M (1998) OFAHIR: On-the-fly automatic authoring of hypertexts for information retrieval. In: Spaccapietra S and Maryanski F, eds. Data Mining and Reverse Engineering: Searching for Semantics, IFIP. Chapman and Hall, pp. 269–300.

Agosti M, Colotti R and Gradenigo G (1991) A two-level hypertext retrieval model for legal data. In: Bookstein A, Chiaramella Y, Salton G and Raghavan V, eds. Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR). Chicago, pp. 316–325.

Agosti M and Crestani F (1993) A methodology for the automatic construction of a hypertext for information retrieval. In: Proceedings of the ACM Symposium on Applied Computing. Indianapolis, USA, pp. 745–753.

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 & Management, 32(4):459–476.

Agosti M, Crestani F and Melucci M (1997) On the use of information retrieval techniques for the automatic construction of hypertexts. Information Processing & Management, 33(2):133–144.

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

Blustein J, Webber R and Tague-Sutcliffe J (1997) Methods for evaluating the quality of hypertext links. Information Processing & Management, 33(2):255–271.

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

Crestani F and Melucci M (1998) A case study of automatic authoring: From a textbook to a hyper-textbook. Data and Knowledge Engineering, 27(1):1–30.

Croft W and Harper D (1979) Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35:285–295.

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

Efthimiadis E (1996) Query expansion. In: Williams M, ed. Annual Review of Information Science and Technology (ARIST), Vol. 31. Information Today for the American Society for Information Science, Medford, NJ, USA, chap. 4, pp. 121–185.

Fox E (1983) Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts. Technical Report TR83-561, Cornell University, Computer Science Department.

Furner J, Ellis D and Willett P (1996) The representation and comparison of hypertext structures using graphs. In: Agosti M and Smeaton A, eds. Information Retrieval and Hypertext. Kluwer Academic, chap. 4, pp. 75–96.

Griffitths A, Luckhurts H and Willett P (1986), Using inter-document similarity information in document retrieval systems. Journal of the American Society for Information Science, 37:3–11.

Harman D (1992) Relevance feedback and other query modification techniques. In: Frakes W and Baeza-Yates R, eds. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, NJ, USA, chap. 11.

Hersh W, Buckley C, Leone T and Hickam D (1994) OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR). Dublin, Ireland, pp. 192–201.

Jardine N and van Rijsbergen C (1971) The use of hierarchical clustering in information retrieval. Information Storage and Retrieval, 7:217–240.

Kwok K (1996), A new method of weighting query terms for ad-hoc retrieval. In: Frei H, Harman D, Schäuble P and Wilkinson R, eds. Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR). New York, pp. 187–196.

Robertson S and Sparck Jones K (1976) Relevance weighting of search terms. Journal of the American Society for Information Science, 27:129–146.

Rocchio JJ (1971) Relevance feedback in information retrieval. In: Salton G, ed. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, NJ, chap. 14, pp. 313–323.

Salton G and Buckley C (1990) Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4):288–297.

Salton G and McGill M (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Salton G, Singhal A, Mitra M and Buckley C (1997) Automatic text structuring and summarization. Information Processing & Management, 33(2):193–207.

Savoy J (1997) Ranking schemes in hybrid Boolean systems: A new approach. Journal of the American Society for Information Science, 48(3):235–253.

Smeaton A (1995) Building hypertext under the influence of topology metrics. In: Proceedings of IWHD Conference. Montpellier.

Smeaton A and Morrissey P (1995) Experiments on the automatic construction of hypertext from text. Technical Report, Dublin City University, School of Computer Applications, Ireland, Working Paper: CA-0295.

Sparck Jones K (1971) Automatic Keyword Classification. Butterworths.

Sparck Jones K (1981) Information Retrieval Experiments. Butterworths.

Sparck Jones K and Willett P (1997) Readings in Information Retrieval. Morgan Kaufmann, San Francisco, CA, USA.

Tague-Sutcliffe J (1992) The pragmatics of information retrieval experimentation, revisited. Information Processing & Management, 28(4):467–490.

Thistlewaite P (1995) Automatic construction of open webs using derived link patterns. In: Agosti M and Allan J, eds. Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR). Seattle, WA.

van Rijsbergen C (1979) Information Retrieval, 2nd ed. Butterworths, London.

van Rijsbergen C and Croft W (1975) Document clustering: An evaluation of some experiments with the Cranfield 1400 collection. Information Processing & Management, 11(5/7):171–182.

Voorhees EM (1985) The cluster hypothesis revisited. Technical Report TR85-658, Computer Science Department, Cornell University.

Willett P (1988) Recent trends in hierarchic document clustering: A critical review. Information Processing & Management, 24(5):577–597.