Enhancing Retrieval with Hyperlinks: A General Model Based on Propositional Argumentation Systems

Justin Picard
MediaSec Technologies, 10 Weybosset Street, Suite 501, Providence, RI 02906.
E-mail: jpicard@mediasec.com

Jacques Savoy
Institut Interfacultaire d'Informatique, Pierre-a-Mazel 7, 2000 Neuchâtel, Switzerland.
E-mail: jacques.savoy@unine.ch

Fast, effective, and adaptable techniques are needed to automatically organize and retrieve information on the ever-increasing World Wide Web. In that respect, different strategies have been suggested to take hypertext links into account. For example, hyperlinks have been used to (1) enhance document representation, (2) improve document ranking by propagating document score, (3) provide an indicator of popularity, and (4) find hubs and authorities for a given topic. Although the TREC experiments have not demonstrated the usefulness of hyperlinks for retrieval, the hypertext structure is nevertheless an essential aspect of the Web, and as such, should not be ignored. The development of abstract models of the IR task was a key factor to the improvement of search engines. However, at this time conceptual tools for modeling the hypertext retrieval task are lacking, making it difficult to compare, improve, and reason on the existing techniques. This article proposes a general model for using hyperlinks based on Probabilistic Argumentation Systems, in which each of the above-mentioned techniques can be stated. This model will allow to discover some inconsistencies in the mentioned techniques, and to take a higher level and systematic approach for using hyperlinks for retrieval.

Introduction

Although many hopes have been placed in the use of hyperlinks for improving information retrieval, their potential impact on retrieval is still a debated question. Conceptual models have been essential to the development of information retrieval as a science. However, no serious attempt has been made yet to model hypertext retrieval. This article proposes to use the framework of Probabilistic Argumentation Systems to represent and use the various types of uncertain knowledge that can be induced from hyperlinks.

Hypertext Links for Enhancing Retrieval: A False Promise?

The difficulties with searching and organizing the Web have led researchers to investigate other sources of knowledge. Recently, attention has focused on one of them: the few billion hypertext links which "glue" the Internet together. The Web would after all not exist without these links, which are the paths that lead to information. Indeed, browsing is for many users the usual way to find "nearby" information, and as quoted in (Marchiori, 1997, p. 265):

> The power of the Web resides in its capability of redirecting the information flow via hyperlinks, so it should appear natural that in order to evaluate the information content of a Web object, the Web structure has to be carefully analyzed.

Hypertext links have been used in different ways, reflecting different assumptions on the type of information they may contain. For example, different authors have studied the impact of incorporating hypertext links to improve an initial ranking of documents for a given query, to produce a better ranking (Frei & Stieger, 1995; Savoy, 1994). In the new context of the Web, they can been used to compute an estimate of a Web page's popularity, assuming that the preferences of users is reflected in the hypertext structure (Brin & Page, 1998). Finally, they can be helpful to organize the Web by category or topic, under the following assumption: for a given topic, important authorities are cited by many important hubs, and important hubs (also called fan pages) cite the important authorities (Bharat & Henzinger, 1998; Chakrabati, Berg, & Don, 1999; Kleinberg, 1998).

The first experiments on the Web using hyperlinks to retrieve or organize information on the Web have yielded significant, if not very significant, improvement over baseline (Bharat & Henzinger, 1998; Brin & Page, 1998; Chakrabati et al., 1999; Kleinberg, 1998; Marchiori, 1997). However, in the only experiments with hypertext collections respecting the standards of IR, namely the TREC-8 and TREC-9 Web track experiments, hypertext links appeared as unpredictable, yielding marginal improvement (Hawking, 2001; Savoy & Picard, 2000).

Leading to excellent results according to some authors, but to marginal improvement when processed in the more rigorous setting of the TREC experiments, the question of the impact of hyperlinks on retrieval effectiveness is a controversial issue. It is not our opinion that, because of their poor results in the TREC experiments, hypertext links are overrated and should be dismissed. However, we think that until now, the hypertext retrieval problem has been essentially studied from an empirical perspective, and it is necessary to develop conceptual tools that would allow considering hypertext retrieval methods from a more abstract perspective.

Modeling Hypertext Retrieval

Even if information retrieval has a very strong empirical tradition, the importance taken by the theoretical work in modeling in a computational way the retrieval process has been essential in the development and maturation of this science. However, although hypertext or citations links have been sometimes integrated in certain models of IR (Fuhr, 1995; Roelleke, Lalmas, & Fahr, 2001; Turtle & Croft, 1991), no general model of hypertext retrieval has been proposed yet, in which each of the different techniques proposed so far could be described.

A model of hypertext retrieval would be useful for several reasons (Sebastiani, 1998):

1. Models are abstractions of the retrieval process, independent of the specific architecture chosen for storing the data, retrieving documents, or acquiring and processing the request. Abstraction leading to generalization, a model would provide a theoretical framework for thinking the hypertext retrieval task.
2. Models provide useful guidelines for developing an operational retrieval system. For example, theoretical arguments can be used to justify that a retrieval system should be built this way rather than that way.
3. Models can also be useful to compare the characteristics of different retrieval approaches in a general way (Bruza & Huibers, 1994; Nie, 1989; Sebastiani, 1998; Turtle & Croft, 1992), reducing the number of experiments and eliminating options which are theoretically unjustified.
4. Furthermore, the intellectual effort needed to build a model leads to put a finger on the underlying assumptions, question them, and possibly replace them by a better set of assumptions.
5. Finally, a model is most of the time based on a well-established theoretical framework. This framework often comes with a range of well-known techniques, which can provide reliable tools for the task at hand.

Different approaches have been taken for modeling the hypertext structure. For example, graph-theoretical approaches can be taken for understanding the connectivity of the Net (Albert, Jeong, & Barabasi, 1999). Although a model of the Web connectivity would certainly be useful for IR purposes, in the first place we are seeking a model to integrate hypertext links in the IR process in a way flexible enough to accommodate all existing techniques, and most desirably the upcoming ones. We are thus focused on developing a conceptual framework that allows to capture the knowledge provided by hyperlinks, not one that allows to describe the connectivity of hypertext collections. It seems then that a connectionist or graph approach would not be appropriate for our goal, because it would not provide the adequate tools to model the interpretation that is assigned to hyperlinks.

Probabilistic models may seem appropriate because they can capture the inherent uncertainty of the IR process, in this case the uncertainty in the knowledge that can be induced from hyperlinks. However, with probabilistic models it is very difficult to capture complex relationships between variables. Probabilistic dependencies can be captured with inference networks, but it was recognized that the hypertext structure cannot be reflected in the inference network, one reason being that potential loops cannot be handled properly (Croft & Turtle, 1993). The following quote of Robertson illustrates the limits of probabilistic models (Robertson & Walker, 1994, p. 232):

> One problem with the formal model approach is that it is often very difficult to take into account the wide variety of variables that are thought or known to influence retrieval. The difficulty arises either because there is no known basis for a model containing such variables, or because any such model may simply be too complex to give a usable exact formula.

Despite the limits of a probabilistic modelling, we do believe that a systematic empirical approach (such as the one taken in Greiff, 1998) to estimate the weight of evidence brought by hypertext links on document relevance would be highly appropriate. However, this article is not about estimating uncertainty, but about representing the knowledge that can be extracted from the hypertext structure. The model we propose comes from the following observation: in each of the proposed techniques, the way hyperlinks are processed is inspired from certain assumptions on the knowledge that can be derived from them. For example, when propagating document scores in the hypertext structure, hyperlinks can be interpreted in the following way: "if a document is cited by a relevant document, then it is possibly relevant itself." Also, when hyperlinks are used to find "popular" Web pages, they can be interpreted in the

following way: "if a document is cited by a popular document, then it is possibly relevant itself." These observations and others will be further justified in the rest of this article.

This type of knowledge can easily be captured by propositional logic, at least if some measure or representation of uncertainty is associated. Probabilistic argumentation systems is a technique for dealing with uncertain knowledge, by integrating propositional logic with probability theory. Propositional logic may not have the flexibility and expressiveness of other logics that have been used to represent and process hypertext knowledge (Fuhr, 1995; Roelleke et al., 2001). But one of the main strengths of Probabilistic Argumentation Systems is its ability to represent uncertainty explicitly. This explicit representation of uncertain inferences is very convenient to understand the inference processes involved in each hypertext retrieval techniques.

In this article we will show how different techniques to handle hypertext links can be described in the unifying framework of Probabilistic Argumentation Systems (PAS). The description on PAS will be short and nontechnical, the focus of this article being on the logical modeling of hypertextual knowledge. For a more technical overview of PAS, the interested reader should consult Haenni, Kohlas, and Lehmann (2000, and references therein). The framework of PAS will allow comparisons of the methods, highlighting some previously unseen weaknesses and leading to new methods for using hyperlinks. However, this article will not present experimental results. These have already been shown in our previous articles (Picard, 1998; Savoy & Picard, 2000; Savoy & Rasolofo, 2001), and the goal of this article is to emphasize PAS as a possible, convenient model to reason on the properties of hypertext retrieval techniques.

Outline of This Paper

The next section introduces the reader to probabilistic argumentation systems, which have already been applied to information retrieval in hypertext (Picard, 1998). Then we will describe techniques to handle hyperlinks to improve document ranking, estimate the popularity of a Web page, and extract the most important hubs and authorities related to a given topic. We will see how probabilistic argumentation systems can be used to deal with the knowledge induced by the hypertext structure for the same purposes. We then make a comparison of the techniques according to their PAS modeling, then proposes some possible extensions. Finally, we will conclude.

Probabilistic Argumentation System

Propositional logic is one of the simplest and most convenient ways of encoding knowledge. An apparent drawback is that propositional logic seems to be unsuitable for taking into account uncertainty. However, uncertainty can be handled rather easily by adjoining particular propositions called *assumptions.* Assumptions are propositions that state the unknown conditions or circumstances upon which the facts and rules depend. If an assumption is known to be true, then the fact or rule that depends on it holds. Otherwise, nothing can be deduced from this fact or rule.

For example, let proposition R_1 denote "document d_1 is relevant to the information need." Proposition R_1 can be either true or false. There might be some uncertainty associated to this fact, for example, it may depend on the reliability of the search engine that has retrieved it. By using an assumption a_1 denoting the uncertain conditions under which the fact R_1 holds, the uncertainty can be captured by: $a_1 \rightarrow R_1$.[1] Similarly, uncertainty in rules can also be captured by assumptions. For example, suppose that documents d_1 and d_2 are concern similar subjects such that in some cases, they are both relevant to the same information need. These conditions that do not always apply can be captured by the following: $l_{12} \rightarrow (R_1 \rightarrow R_2)$, where R_2 means "document d_2 is relevant," and l_{12} denotes the uncertain circumstances under which the rule $R_1 \rightarrow R_2$ applies. We will, in general, prefer the following equivalent notation for uncertain rules: $R_1 \wedge l_{12} \rightarrow R_2$.

Most applications also require a numerical assessment of uncertainty. The numerical aspect of uncertainty is obtained by assigning probabilities to assumptions. For example, if for the uncertain rule $R_1 \wedge l_{12} \rightarrow R_2$, the condition l_{12} is known to hold with probability 0.3, then we may write: $p(l_{12}) = 0.3$. Note that this is conceptually different from assigning a probability to the whole logical sentence ($p(R_1 \rightarrow R_2) = 0.3$), as is done in other frameworks for integrating uncertainty with logic.

Given a knowledge base composed of uncertain facts, rules, or conditions modeled with logical formulas containing assumptions, we are interested in finding which symbolic *arguments* support or discard a given hypothesis h. A symbolic argument is a conjunction of literals of assumptions which, if added to the knowledge base, makes the hypothesis true. We will then compute the *symbolic support* of h given by the knowledge base ξ, denoted $sp(h, \xi)$, which contains the disjunction of all symbolic arguments that allow to derive h if added to the knowledge base. We may also want to evaluate the reliability of the support given by arguments, using probabilities assigned to the assumptions. We will then compute the *degree of support* $dsp(h, \xi) = p(sp(h, \xi))$, the probability that the hypothesis h is supported by the knowledge base ξ.

An Example

For example, consider the following knowledge base:

$$\xi = (a_1 \rightarrow R_1) \wedge (a_2 \rightarrow R_2) \wedge (R_1 \wedge l_{12} \rightarrow R_2) \quad (1)$$

Remark that a knowledge base can always be represented as a conjunction of rules, facts, and more generally clauses.

[1] For reading commodity, assumptions will be denoted by lowercase letters, and other propositions by capital letters.

We are interested in finding the arguments for hypothesis R_2 given by ξ. It is easily seen that a_2 is an argument for R_2, because $(a_2 \rightarrow R_2) \wedge a_2 \models R_2$. The same way, $(a_1 \wedge l_{12})$ is another argument for R_2. Thus, the symbolic support given by the knowledge base for R_2 is computed in the following way:

$$sp(R_2, \xi) = a_2 \vee (a_1 \wedge l_{12}) \tag{2}$$

The following probabilities are assigned to the assumptions: $p(a_1) = 0.4$, $p(a_2) = 0.25$, $p(l_{12}) = 0.3$. What is the probability that the support holds? To make an exact computation, independence assumptions must be made, e.g., $p(a_1 \wedge a_2) = p(a_1) \cdot p(a_2)$, $p(a_1 \wedge \neg a_2) = p(a_1) \cdot (1 - p(a_2))$, etc. The degree of support of R_2 given by the knowledge base, $dsp(R_2, \xi)$, is:

$$dsp(R_2, \xi) = p(sp(R_2, \xi)) \tag{3}$$

$$= p(a_2 \vee (a_1 \wedge l_{12})) \tag{4}$$

$$= p(a_2 \vee (a_1 \wedge l_{12} \wedge \neg a_2)) \tag{5}$$

$$= p(a_2) + p(a_1) \cdot p(l_{12}) \cdot (1 - p(a_2)) \tag{6}$$

$$= 0.34 \tag{7}$$

The passage from Equation 4 to 5 in the previous equations comes from the logical equivalence: $A \vee B = A \vee (B \wedge \neg A)$. The next sections will show how the notions on PAS presented here can be applied to deal with the knowledge induced by the hypertext structure.

Using Hyperlinks to Modify Document Score and Rank

Spreading Activation in Hypertext

The implicit reasoning made in the spreading activation (SA) technique is the following: a link from a document d_1 to a document d_2 is evidence that their content is similar or related, such that if d_1 is relevant to a given request, d_2 may also be relevant. From an initial ranking of documents produced by a search engine, the hyperlinks can be used to improve the ranks of the documents linked to the best ranked documents. For example, if d_2 is linked to d_1 which is ranked first, then d_2 should be placed at a better rank.

The general scheme works as follows. The retrieval engine computes an initial retrieval status value (RSV) or score for each document based on its similarity with the query q. The RSV of document d is then updated by adding a fraction of the RSV of its m neighbors through a certain number of cycles. The neighbors can linked by incoming but also outgoing links. Suppose that document d has neighbors d_1 to d_m. The RSV of d at cycle $i + 1$ is computed by the following:

$$\mathrm{RSV}(d^0) = \mathrm{score}(d, q) \tag{8}$$

$$\mathrm{RSV}(d^{i+1}) = \mathrm{RSV}(d^i) + \sum_{j=1}^{m} \lambda_j \cdot \mathrm{RSV}(d_j^i) \tag{9}$$

The parameter λ_j can be seen as the degree of certainty regarding the evidence provided by the link from d_j to d. It can be a fixed value according to the link type,[2] or may vary according to a measure of similarity between the documents and the query (Savoy, 1997). We may also repeat this propagation scheme through a certain number of cycles under the assumption that "friends of my friends are my friend." However, the number of cycles is often limited to one: more than one cycle is usually harmful to retrieval effectiveness (Savoy, 1997).

Several problems can be found with SA approach: there is no theoretical background guiding the choice of the number of cycles c or the value of the parameter λ_j. Moreover, evidence may propagate more than once if there are cycles in the network. That does not mean that spreading activation is not a suitable technique for hypertext retrieval: it was demonstrated in Crestani and Lee (2000) that *constrained* spreading activation can significantly improve retrieval results of Web search engines. As can be guessed, in constrained spreading activation, different types of constraints limit the spreading of retrieval status values.

Improving Document Ranking with PAS

This technique can be modeled within the PAS framework, also to improve document ranking using the hypertext structure. But in the PAS modeling, instead of propagating document scores, we will seek all symbolic arguments supporting the relevance of a document. In a second phase, probabilities are assigned to the assumptions and the degree of support given by the arguments is computed. Finally, documents are returned to the user by decreasing degree of support.

For each document d_i, let us denote proposition R_i as: "document d_i is relevant." If a document is retrieved by the retrieval system, this is evidence in favor of that document's relevance. Let assumption a_i denote, "the retrieval system has correctly retrieved document d_i." Then for each document in the collection, we have:

$$a_i \rightarrow R_i \tag{10}$$

For a given query, we may adjust the probability of the assumption $p(a_i)$ to the rank at which d_i is retrieved ($p(a_i|rank)$), and set $p(a_i) = 0$ if d_i is not retrieved. In

[2] We may consider links of various types: hypertext links, citation, nearest neighbor, etc. Moreover, links can be distinguished by their orientation (incoming, outgoing), because this orientation may affect the amount of information about relevance contained in the link.

practice, we could fit a logistic regression on the rank for a set of training queries (Picard, 1998).

In a second step, we want to use the hypertext structure to improve this initial ranking. For each link from d_i to d_j, we induce the knowledge that, under some condition l_{ij}, the relevance of d_i implies the relevance of d_j. The assumption l_{ij} may denote the conditions under which the link implies relevance in the present context. We have then:

$$R_i \wedge l_{ij} \rightarrow R_j \tag{11}$$

This rule can be read as: "If document d_i is relevant, then, under some condition l_{ij} (that the link from d_i to d_j implies d_j's relevance), d_j is also relevant."

A hyperlink can imply a semantic relationship that takes place in the two directions. Hyperlinks can then imply relevance in the backward direction:

$$R_j \wedge l_{ji} \rightarrow R_i \tag{12}$$

Equations 11 and 12 correspond respectively to a modelling of forward spreading activation and backward spreading activation. Let us denote ξ^F (F for "forward") as the body of knowledge generated from Equations 10 and 11. Also, let us denote ξ^B (B for "backward") as the body of knowledge generated from Equations 10 and 12. Then, when evaluating a certain hypothesis such as R_i, we are free to compute the support from either one of the knowledge bases, or both, i.e.: $sp(R_i, \xi^F)$, $sp(R_i, \xi^B)$ and $sp(R_i, \xi^F \wedge \xi^B)$, where $\xi^F \wedge \xi^B$ is the conjunction of the two knowledge bases. We may then find the arguments supporting the relevance of each document.

As an example, take a collection containing documents d_1, d_2, d_3. There are links from d_2 to d_1 and from d_3 to d_1. Considering links in the forward direction, the following knowledge base is generated:

$$\xi^B = (a_1 \rightarrow R_1) \wedge (a_2 \rightarrow R_2) \wedge (a_3 \rightarrow R_3) \wedge (R_2 \wedge l_{12} \rightarrow R_1) \wedge (R_3 \wedge l_{31} \rightarrow R_1) \tag{13}$$

We find for the support of R_1 given by the knowledge base ξ^B:

$$sp(R_1, \xi^B) = a_1 \vee (a_2 \wedge l_{21}) \vee (a_3 \wedge l_{31}) \tag{14}$$

Here, d_1 has three symbolic arguments. For a real query, one may want a numerical evaluation. The degree of support of R_1 is:

$$\begin{aligned} dsp(R_1, \xi^B) &= p(sp(R_1, \xi)) = p(a_1) \\ &+ p(a_2 \wedge l_{21} \wedge \neg a_1) + p(a_3 \wedge l_{31} \wedge \neg a_1 \wedge \neg(a_2 \wedge l_{21})) \\ &= p(a_1) + p(a_2) \cdot p(l_{21}) \cdot (1 - p(a_1)) + p(a_3) \cdot p(l_{31}) \\ &\quad \cdot (1 - p(a_1)) \cdot (1 - p(a_2) \cdot p(l_{21})) \end{aligned}$$

For a given query, one needs to give values to the $p(a_i)$s according to the rank of d_i, and probabilities for the links $p(l_{ij})$. It is interesting to notice that the hypertext structure, interpreted logically, has been integrated in the computing formulas for the degree of support of each document. This way, the computations can be done very quickly. More details on the implementation of the PAS model for spreading activation can be found in Picard (1998) and Savoy and Picard (2000).

As in the spreading activation technique, the PAS implementation of the model uses a computing formula. However, the PAS model demonstrates several interesting features:

1. Loops or cycles in the hypertext structure are naturally handled and are not pathological cases.
2. Variables are meaningful, as they correspond to the probability that a link implies relevance ($p(l_{ij})$) and the probability that a document is relevant given its rank ($p(a_i)$).
3. Evidence is a propagated in a sound way, following the rules of propositional logic.
4. It is possible to combine forward and backward interpretation of the rules in the same knowledge base, while this would create loops with the spreading activation technique.

Estimating the Popularity of a Web Page

PageRank

The PageRank algorithm (Brin & Page, 1998) considers that users have an absolute preference among Web pages: it assumes that the more a Web page is visited, the more it is appreciated by the users. To measure this popularity, a reasonable assumption is that the preference of users is reflected in the hypertext structure: a link toward a Web page is often an indication that this page is acknowledged by the author as a good source of information. A simple way to implement this idea would be to count the number of times a Web page is cited. Microsoft's home page, surely one of the most visited page on the Web, is cited more than 23 million times in Altavista's index (probably much more in reality). However, each link should not be treated equally, because its impact also depends on the popularity of the parent node: a page cited only a few times but which is in Yahoo!'s index would certainly be quite visited. Thus, the popularity of a page also depends on the popularity of the pages that cite it.

Such a popularity measure is used in the Google search engine to boost the scores of the documents, independently of the query. This algorithm is criticized because it biases the access to information (Lawrence & Giles, 1999). The "perverse" effect of PageRank is that it will push popular pages to get even more popular, and new or unknown (unlinked) Web pages to stay unknown. As said in Marchiori (1997), "visibility is likely to be a synonym of popularity, which is completely different than quality, and thus

using it to gain higher score is a rather poor choice." To our advice, the frequency at which a page is visited by all the users of the Web is not necessarily an indicator of its relevance to a user who has its own preferences, cultural background, etc. As we shall now see, the PAS modeling offers a clean way to take account of these a priori user preferences.

Measuring Popularity with PAS

Suppose that for each user it is possible to compute some personalized "popularity" measure. For example, each user may define a profile (e.g., a set of keywords, a set of Web pages defined by a bookmarks list), such that, for each page on the Web, we can assign a probability $p(a_i)$ based on its similarity with the user profile. We would like to refine this *prior* knowledge by taking account of the hypertext structure. Let us define P_i as "document d_i corresponds to the user's interest." Then for each document, there is some condition d_i under which d_i corresponds to the user's interest that is denoted as:

$$a_i \rightarrow P_i \tag{15}$$

The probability $p(a_i)$ can be initially computed bases on the user profile, and then updated by keeping track of the pages visited. For each link from d_i to d_j, we induce that, under some condition l_{ij}, the relevance of d_i implies the relevance of d_j to the user's interest. We have then:

$$P_i \wedge l_{ij} \rightarrow P_j \tag{16}$$

In this model, each user will have the same symbolic arguments supporting the popularity of each document. If an equal probability $p(a_i)$ is assigned to each document, it is assumed that the user has no preference, and this case corresponds to the PageRank model. However, if the user gives some hints allowing to compute personal a priori probabilities $p(a_i)$ or eventually link probabilities (e.g., by inspecting user bookmarks lists), it will be possible to have a personal ranking for this user.

Finding Hubs and Authorities

Kleinberg's Algorithm

In many cases, the user does not know what exactly he/she is looking for, and is rather interested in having good starting points for browsing. Given a general topic sufficiently represented on the Web, it is possible to distinguish two types of potentially relevant pages: authorities and hubs. Authorities are pages containing high quality and exhaustive information on a topic, and hubs are pages containing links to the authorities, thus giving access to the relevant information. The Web is rich in central pages, fan sites, and other classifications of resources, and those can be very helpful for automatic classification of information.

How can we find hubs and authorities? The assumption made by Kleinberg (1998) is that a good authority is a page that has links from many good hubs, and a good hub is a page that has links towards many good authorities. The algorithm has some similarity with PageRank model in that the quality of a page depends recursively on the quality of the neighbors, although here the links are followed in both directions. The idea is implemented in the HITS algorithm as follow (Kleinberg, 1998): first, a root set is extracted, containing the most likely relevant pages found with a search engine in response to a given query (e.g., 200 documents). This root set is expanded with all documents that point to or are pointed by these pages, to form the base set in which authorities and hubs will be found. Then the connectivity of this base set is used as follows to find the best hubs and authorities. For each document d_p in the base set, a hub score h_p and a authority score a_p are computed. Both initial scores are set to 1. Then the hub and authority scores are updated iteratively by, respectively, the sum of authority scores of pages cited by d_p, and the sum of hub scores of the pages citing d_p. The updating equations are:

$$h_p = \sum_{d_p \rightarrow d_i} a_i,\ a_p = \sum_{d_i \rightarrow d_p} h_i \tag{17}$$

where $d_p \rightarrow d_i$ means that there is a link from d_p to d_i. It can be shown that both scores will converge if they are normalized after each iteration. The exact scores are not so important, because the user is presented with a ranked list of hubs and authorities.

It is argued that the algorithm has an "objective" justification because it finds some intrinsic properties of a set of linked pages (Chakrabati et al., 1999; Kleinberg, 1998). However, some aspects of the algorithm are arbitrary: for example, a base set has to be chosen for a given topic, from which the most important hubs and authorities will be selected. Although it is argued in Kleinberg (1998) that the method is robust (i.e., gives similar results) for different base sets, the choice of a particular base set is nonetheless purely heuristic. Another questionable aspect of the algorithm is that the initial ranking of documents is not used as prior evidence (i.e., all documents in the base set are treated equally), while it is likely that an initially better ranked document has more chances to be relevant, and certainly more chances to be a good hub or authority.

A Model for Computing Hub and Authority Scores

For document d_i, proposition H_i denotes "document d_i is a hub" and A_i denotes "document d_i is an authority," independently of any particular request. We consider that there is initial evidence h_i that R_i is a good hub, and a_i that it is a good authority. This evidence can be given by an initial ranking of documents:

$$h_i \rightarrow H_i,\ a_i \rightarrow A_i \tag{18}$$

As in Kleinberg's algorithm, we make the assumption that if a document d_i is cited by a good hub d_j, then this is evidence that d_i is a good authority. We have then:

$$H_j \wedge f_{ji} \rightarrow A_i \tag{19}$$

Similarly, if a good authority d_i is cited by a document d_j, then this is evidence that the d_j is a good hub:

$$A_i \wedge g_{ij} \rightarrow H_j \tag{20}$$

For each hyperlink from d_i to d_j, there will be two rules generated: $(H_j \wedge f_{ji} \rightarrow A_i)$, $(A_i \wedge g_{ij} \rightarrow H_j)$. From this knowledge base ξ, one can compute for each document d_i, the symbolic support from ξ that it is a good hub and a good authority, $sp(H_i, \xi)$ and $sp(A_i, \xi)$. Then, for a given topic, different probabilities are assigned to the assumptions $p(d_i)$ and $p(h_i)$, and eventually to the assumptions f_{ij} and g_{ij}, which can be fixed or depend on some similarity value with the topic. The numerical degrees of support $dsp(H_i, \xi)$ and $dsp(A_i, \xi)$ will be the hub and authority scores of document d_i for this topic. Note that compared with Kleinberg's algorithm, there is no need to determine a base set, which is here the same for all topics. For different topics, only the assigned probabilities will change.

Discussion

Comparison of Techniques

In the previous sections we have seen different ways in which hypertext links can be interpreted as uncertain knowledge, which is then converted into propositional sentences using the PAS framework. Table 1 summarizes the PAS modelling of the different techniques.

There is a striking similarity between the PAS modeling of the popularity measure and of the forward ranking. Indeed, by replacing the R_is by P_is, we would obtain the same knowledge base, and this way the same symbolic support for each document. This means that although they seem different in their spirit, the methods are, in fact, based on the same interpretation of the knowledge contained in the hyperlinks. Because the popularity measure technique is equivalent to the forward ranking, we can conclude that the former is equivalent to a specific case of the latter, in which hyperlinks can also induce rule in the backward direction.

TABLE 1. Comparison of the different techniques

Technique	Propositions	A priori evidence	Link (d_i, d_j)
Ranking (forward)	R_i	$a_i \rightarrow R_i$	$R_i \wedge l_{ij} \rightarrow R_j$
Ranking (backward)	R_i	$a_i \rightarrow R_i$	$R_j \wedge l_{ji} \rightarrow R_i$
Popularity	P_i	$a_i \rightarrow P_i$	$P_i \wedge l_{ij} \rightarrow P_j$
Hubs, authorities	A_i,H_i	$a_i \rightarrow A_i$	$H_i \wedge f_{ij} \rightarrow A_i$
		$h_i \rightarrow H_i$	$A_j \wedge g_{ji} \rightarrow H_i$

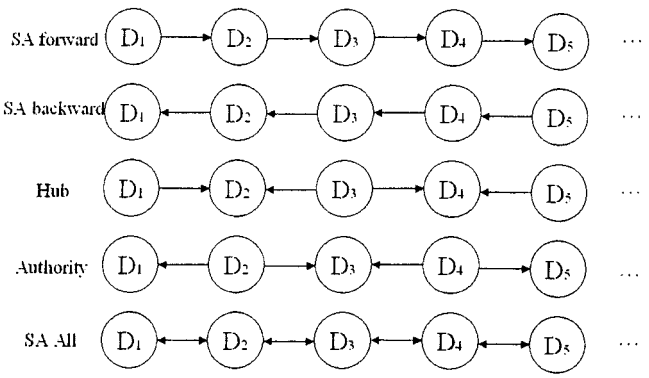

FIG. 1. Sequences of hyperlinks needed to produce an argument.

Making a comparison between the ranking and the hub and authorities modeling is not as straightforward. The hub and authorities modeling seems to make a more subtle use of the hyperlinks, because two propositional symbols A_i and H_i are used for each document, and are linked through uncertain rules in a symmetric ways. To allow for a comparison, Figure 1 can be useful. In this figure, documents d_1 and d_5 are indirectly hyperlinked through d_2, d_3, and d_4. The graphic represents, for different possible hypothesis concerning d_5, i.e., R_5, A_5, and H_5, the sequences of hyperlink direction that is required for having an inference chain from d_1 to d_5. For example, for the forward ranking technique (SA forward on the figure), the required sequence of links is $d_1 \rightarrow d_2 \rightarrow d_3 \rightarrow d_4 \rightarrow d_5$, which will lead to an argument $a_1 \wedge l_{12} \wedge l_{23} \wedge l_{34} \wedge l_{45}$ for R_5. Any other sequence of link directions would not lead to an argument.

It is straightforward to see how an argument is produced for the ranking techniques. Note that if the forward and backward knowledge base are combined (SA All), any sequence of link direction leads to an argument. This technique can be considered as the most flexible on the hyperlink direction. For the hub and authorities modelling, one can see that to produce an argument, link directions have to alternate: forward-backward-forward-backward for d_1 to produce a hub argument for d_5 (i.e., support proposition H_5, and reverse way to produce an authority argument.

What conclusion can we drawn from this graphical comparison? Well, the spirit of a technique for handling hyperlinks can be represented, graphically, as the sequence of link directions that "triggers" an argument for two indirectly linked documents. This conceptual tool can be useful as an objective comparison of two different techniques, or can lead to develop new techniques that follow certain desired properties. Indeed, it is possible to generate uncertain rules only if specific sequences of hyperlinks are detected, for example, loops, cocitation (two documents citing a third one), and so on.

Extensions

Many extensions are possible to the PAS modeling of hypertext retrieval techniques. Generally, these extensions

can be seen as other bodies of knowledge (other propositions, other assumptions, and other rules) combined with the existing one. In the PAS framework, different knowledge bases can be integrated as long as they do not contradict each other.

One interesting body of knowledge that could be added relates to the "quality" of the links. Indeed, depending on the context, the same hypertext link can indicate appropriate semantic relationships, or to the contrary, be misleading. As has been shown by different authors, taking this context into account can lead to a significant performance increase (Bharat & Henzinger, 1998; Chakrabati et al., 1999; Frei & Stieger, 1995). For example, it is possible to measure a similarity between the hyperlink anchor and the query, and weight the score propagated through the hyperlink accordingly (Chakrabati et al., 1999). Such measure of the link quality can be incorporated numerically in the PAS framework by adjusting the link probabilities. The probability l_{ij} that a link indicates relevance can be made higher when the two documents are more similar, or when the query is more similar to the hyperlink anchor (Picard, 2000).

However, in this article we focus on the problem of representing uncertain knowledge. For this case, we would like to capture the knowledge that the "validity" of a hyperlink depends on the context, more precisely on the query formulated by the user, or the topic for which the hubs and authorities have to be extracted. Assume we take propositional symbols $T_1, \ldots, T_N$ to represent the context. These symbols may refer to the query terms, or to the topic. Typically, a topic or query can be represented as a conjunction of terms, for example, $T_1 \wedge T_2$. Then, these propositions can imply the truth of link assumptions. For example, in the ranking interpretation of Hyperlinks, suppose we have a link from d_i to d_j and the following rule:

$$T_1 \rightarrow l_{ij} \tag{21}$$

Then, if proposition T_1 is true (e.g., the corresponding term is in the query), the rule $D_i \wedge l_i j e D_j$ is equivalent to $D_i \rightarrow D_j$, because l_{ij} becomes true. On the other hand, another query term may render the same hyperlink invalid, for example, $T_2 \rightarrow \neg l_{ij}$.

This example can be easily extended to the hub and authorities body of knowledge. Of course, several other extensions are possible. The reader is referred to Picard (2000) for a discussion on the inclusion of different bodies of knowledge in the PAS framework.

Conclusion

Considering information retrieval from a logical viewpoint has brought much enlightening on its underlying mechanisms. The logical approach has led to the creation of meta-models of IR, in which different approaches can be described and analyzed to illustrate if they possess some general properties. In the same vein, Nie made the demonstration that some forms of vector-space model are inconsistent (Nie, 1989). Huibers and Bruza determined a set of axioms concerning information carriers, and demonstrated that Boolean retrieval is superior to some form of coordination match (Bruza & Huibers, 1994). Crestani and van Rijsbergen explored the mechanism of probability transfer in IR, using logical imaging (Crestani & van Rijsbergen, 1995). More recently, Dominich has developed a unified axiomatic foundation for the classical models (Dominich, 2001).

This article goes in the same line. It was shown how various techniques to the use of hyperlinks to search and organize the Web could be described in the framework of logic. Our approach can be summarized as follow: (1) translate the sometimes implicit principles of each technique into explicit assumptions on the knowledge that can be derived from hypertext links, (2) show how this knowledge can be expressed into propositional logic if uncertainty is described with special prepositional symbols, (3) use the theoretical framework of Propositional Argumentation Systems to generate a knowledge base and use it to make inferences. As the long as the essence of a technique to handle hyperlink can be translated as a certain way to induce knowledge from these links, it can be modelled within this theoretical framework.

Translating different hypertext retrieval techniques into the PAS framework as different bodies of knowledge allowed a formal analysis of these approaches on the ground of logic. This modeling shed light on inconsistencies or weaknesses in the implementation of the techniques. Similarities between techniques, such as the equivalence between the popularity measure technique and the forward spreading activation, could be clearly highlighted. On the other hand, techniques very different in their spirit could be compared at an abstract level, as different ways in which sequences of links lead to inferences in the PAS framework.

By modeling hypertext retrieval techniques with logic, we can rely on the strict rules of logic for propagating evidence in the hypertext structure. Having uncertainty represented with symbols (i.e., assumptions) in the knowledge base leaves us free hands for developing tools to assign appropriate values to the associated probabilities. In that view, separating the representation of uncertainty from its assessment is a convenient "divide-and-conquer" way to tackle the problem of building more effective IR systems.

References

Albert, R., Jeong, H., & Barabasi, A. (1999). Diameter of the World-Wide-Web. Science, 401, 130.

Bharat, K., & Henzinger, M. (1998). Improved algorithms for topic distillation in hyperlinked environments. In Proceedings of the international ACM-SIGIR conference, Melbourne, Australia (pp. 104–111).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In Proceedings of the World Wide Web Conference, Brisbane, Australia (pp. 107–117).

Bruza, P., & Huibers, T. (1994). Investigating aboutness axioms using information fields. In Proceedings of the International ACM-SIGIR conference, Dublin, Ireland (pp. 112–121).

Chakrabati, S., der Berg, M.V., & Dom, B. (1999). Focused crawling: A new approach to topic specific resource discovery. In Proceedings of the World Wide Web conference, Toronto, Canada (pp. 545–567).

Crestani, F., & Lee, P. (2000). Searching the web by constrained spreading activation. Information Processing and Management, 36(4), 585–605.

Crestani, F., & van Rijsbergen, C. (1995). Information retrieval by logical imaging. Journal of Documentation, 51(1), 3–17.

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

Dominich, S. (2001). Mathematical foundations of information retrieval. Kluwer Academic Publishers.

Frei, H., & Stieger, D. (1995). The use of semantic links in hypertext information retrieval. Information Processing & Management, 31(1), 1–13.

Fuhr, N. (1995). Probabilistic datalog—A logic for powerful retrieval models. In Proceedings of the international ACM-SIGIR conference, Seattle, Washington (pp. 282–290).

Greiff, W. (1998). A theory of term weighting based on exploratory data analysis. In Proceedings of the international ACM-SIGIR conference, Melbourne, Australia (pp. 11–19).

Haenni, R., Kohlas, J., & Lehmann, N. (2000). Probabilistic argumentation systems. In J. Kohlas & S. Moral (eds.), Handbook of defeasible reasoning and uncertainty management systems, Vol. 5: Algorithms for uncertainty and defeasible reasoning. Dordrecht: Kluwer.

Hawking, D. (2001). Overview of the TREC-9 web task. TREC-9.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In Proceedings of the 9th symposium on discrete algorithms (pp. 668–677).

Lawrence, S., & Giles, C. (1999). Accessibility of information on the web. Nature, 400(8), 107–109.

Marchiori, M. (1997). The quest for correct information on the Web: Hyper search engines. In Proceedings of the World Wide Web conference, Santa Clara, California.

Nie, J. (1989). An information retrieval model based on modal logic. Information Processing & Management, 25(5), 477–491.

Picard, J. (1998). Modeling and combining evidence provided by document relationships using probabilistic argumentation systems. In Proceedings of the international ACM-SIGIR conference on research and development in information retrieval (182–189). Melbourne, Australia.

Picard, J. (2000). Probabilistic argumentation systems applied to information retrieval. PhD thesis, University of Neuchâtel.

Robertson, S., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the International ACM-SIGIR conference, Dublin, Ireland (pp. 232–241).

Roelleke, T., Lalmas, M., & Fuhr, N. (2001). Intelligent retrieval of hypermedia documents. In Intelligent Exploration of the Web. Heidelberg, Germany: Physica-Verlag, to appear.

Savoy, J. (1994). A learning scheme for information retrieval in hypertext. Information Processing & Management, 30(4), 513–533.

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

Savoy, J., & Picard, J. (2000). Report on the TREC-8 experiment: Searching on the Weband in distributed collections. In D. Harman (Ed.), TREC-8. Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD.

Savoy, J., & Rasolofo, Y. (2001). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In D. Harman, (ed.), TREC'9. Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD.

Sebastiani, F. (1998). On the role of logic in information retrieval. Information Processing and Management, 34(1), 1–18.

Turtle, H., & Croft, W. (1991). Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 9(3), 187–222.

Turtle, H., & Croft, W. (1992). A comparison of text retrieval models. The Computer Journal, 35(3), 279–290.