Using Probabilistic Argumentation Systems to Search and Classify Web Sites

Justin Picard and Jacques Savoy *
Institut interfacultaire d'informatique
Pierre-à-Mazel 7
2000 Neuchâtel (Switzerland)

Abstract

As the amount of information stored on the web increases at an amazing pace, it gets harder for search engines to retrieve the needed information. Recently, attention in the web community has focused on the use of hyperlinks to help index and organize information. Several methods have been suggested to take account of the knowledge induced by hyperlinks, with different objectives in mind. In this paper, we focus on three of them: improving document ranking, estimating the popularity of a web page, and extracting the most important hubs and authorities related to a given topic. Using probabilistic argumentation systems, a technique for dealing with uncertain knowledge which integrates propositional logic and probability theory, we show how all these techniques can be modeled in a unified logical framework. This allows comparisons of the different methods for using hyperlinks and illustrates some of their weaknesses based on some experiments.

1 Introduction

Traditional techniques for indexing automatically documents must be reviewed in the context of the Internet. For example, web pages may vary from a few words to megabytes. Moreover, the assumption that the words contained in a document are good indicators of the underlying semantics is not as much verified as in document collections: people who write web pages may have specific objectives in mind, and use tactics to influence the ranking of search engines, called "spamming" or "the search engine persuasion problem" [Mar97]. Finally, the well-known problems of polysemy and synonymy are aggravated on the web: in a click, one can find commercial information, scientific papers, data repositories, propaganda or the web pages of friends.

To compensate the difficulties with indexing and retrieving information, many search engines are also offering hierarchical classifications. These classifications were created and maintained manually. However the cost of such a manually-based organization of information does not make it a viable option at long term because web pages change continuously and new categories or "cyber communities" appear everyday on the web. And as reported in [Mar97], "repositories are now themselves resorting to search engines to keep their database up-to-date".

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

* This research was supported in part by the SNF (Swiss National Science Foundation) under grant 20'55'771.98

1.1 Hypertext links

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 which lead to information. Indeed, browsing is for many users an usual way to find "nearby" information [Mar97].

Recent experiments seem to indicate that hyperlinks can be very valuable in locating or organizing information. They have been suggested: (1) to enhance an initial ranking of documents [Sav94], (2) to compute an estimate of a web page popularity [BL98], or (3) to find the most important hubs and authorities for a given topic [Kle98, Bh98, CdBD99]. It seems that each of these techniques is based on some underlying, sometimes partly implicit hypothesis, on the type of knowledge which can be induced by the presence of a hyperlink between two pages. For example, when they are used to estimate the popularity of a web page, hyperlinks can be interpreted in the following way: "if a document is cited by a popular page, then it is possibly popular itself". This type of knowledge can easily be captured by propositional logic, if some measure of uncertainty is associated.

1.2 Outline of this paper

Probabilistic argumentation systems are a convenient technique for dealing with uncertain knowledge, by combining propositional logic with probability theory. In this paper, we will show how the techniques mentioned in the previous section can be described in this unified framework. This allows comparisons of the methods, highlights some previously unseen weaknesses and leads to new methods using hyperlinks.

Section 2 will introduce the probabilistic argumentation systems, which have already been applied to information retrieval (IR) in hypertext [Pic98]. Sections 3, 4 and 5 will describe techniques to handle hyperlinks in order to improve document ranking, estimate the popularity of a web page, and extract the most important hubs and authorities related to a given request. We will see how probabilistic argumentation systems can be used to deal with the uncertain knowledge induced by the hypertext structure for the same purposes. Section 6 will present some experiments with the model developed for improving an initial ranking of documents and the last section will conclude this paper.

2 Probabilistic argumentation systems

In this section, we make a short introduction to probabilistic argumentation systems (PAS). There is much more about PAS than what is shown here. For a detailed overview of PAS, the reader is referred to [HKL99]. However, this tutorial should be sufficient to understand their application in this paper.

2.1 PAS

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 account of uncertainty. However, uncertainty can be handled rather easily by adjoining particular propositions called **assumptions**. Assumptions are propositions which 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 which depends on it holds. Otherwise, nothing can be deduced from this fact or rule.

For example, let proposition D_1 denote "document d_1 is relevant to the information need". Proposition D_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 which has retrieved it. By using an assumption a_1 denoting the uncertain conditions under which the fact D_1 holds, the uncertainty can be captured by: $a_1 \rightarrow D_1$ [1]. Similarly, uncertainty

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

in rules can also be captured by assumptions. For example, suppose that documents d_1 and d_2 concern similar subjects such that in some cases, they are both relevant to the same information need. These conditions which do not always apply can be captured by the following: $l_{12} \rightarrow (D_1 \rightarrow D_2)$, where D_2 means "document d_2 is relevant", and l_{12} denotes the uncertain circumstances under which the rule $D_1 \rightarrow D_2$ applies. We will in general prefer the following equivalent notation for uncertain rules: $D_1 \wedge l_{12} \rightarrow D_2$.

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 $D_1 \wedge l_{12} \rightarrow D_2$, 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 ($p(D_1 \rightarrow D_2) = 0.3$), 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 which 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.

2.2 An example

For example, consider the following knowledge base:

$$\xi = (a_1 \rightarrow D_1) \wedge (a_2 \rightarrow D_2) \wedge (D_1 \wedge l_{12} \rightarrow D_2) \tag{1}$$

Remark that a knowledge base can always be represented as a conjunction of rules, facts, and more generally clauses. We are interested in finding the arguments for hypothesis D_2 given by ξ. It is easily seen that a_2 is an argument for D_2, because $(a_2 \rightarrow D_2) \wedge a_2 \models D_2$. The same way, $(a_1 \wedge l_{12})$ is another argument for D_2. Thus, the symbolic support given by the knowledge base for D_2 is computed in the following way:

$$sp(D_2, \xi) = a_2 \vee (a_1 \wedge l_{12}) \tag{2}$$

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? In order 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 D_2 given by the knowledge base, $dsp(D_2, \xi)$, is:

$$\begin{aligned}
dsp(D_2, \xi) &= p(sp(D_2, \xi)) && (3)\\
&= p(a_2 \vee (a_1 \wedge l_{12})) && (4)\\
&= p(a_2 \vee (a_1 \wedge l_{12} \wedge \neg a_2)) && (5)\\
&= p(a_2) + p(a_1) \cdot p(l_{12}) \cdot (1 - p(a_2)) && (6)\\
&= 0.34 && (7)
\end{aligned}$$

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

3 Using hyperlinks to modify document score and rank

3.1 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 an 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 enhance 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:

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

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

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 [Sav97]. We may also repeat this propagation scheme through a certain number of cycles under the assumption that ”friends of my friends are my friends”. However, the number of cycles is often limited to one: more than one cycle is usually harmful to retrieval effectiveness [Sav97].

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.

3.2 Improving document ranking with PAS

Based on the PAS approach, we will also attempt to improve document ranking using the hypertext structure. But 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 D_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:

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

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 practice, we fit a logistic regression on the rank for a set of training queries [SP01].

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.

[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.

The assumption l_{ij} may denote the conditions under which the link implies relevance in the present context. We have then:

$$D_i \wedge l_{ij} \rightarrow D_j \tag{11}$$

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. The following knowledge base is generated:

$$\xi = (a_1 \rightarrow D_1) \wedge (a_2 \rightarrow D_2) \wedge (a_3 \rightarrow D_3) \wedge (D_2 \rightarrow D_1) \wedge (D_3 \rightarrow D_1) \tag{12}$$

We find for the support of D_1:

$$sp(D_1, \xi) = a_1 \vee (a_2 \wedge l_{21}) \vee (a_3 \wedge l_{31}) \tag{13}$$

Here, d_1 has three symbolic arguments. For a real query, one may want a numerical evaluation. The degree of support of D_1 is:

$$\begin{aligned} dsp(D_1, \xi) &= p(sp(D_1, \xi)) = p(a_1) + p(a_2 \wedge l_{21} \wedge \neg a_1) + p(a_3 \wedge l_{31} \wedge \neg a_1 \wedge \neg(a_2 \wedge l_{21})) \\ &= p(a_1) + p(a_2) \cdot p(l_{21}) \cdot (1 - p(a_1)) + p(a_3) \cdot p(l_{31}) \cdot (1 - p(a_1)) \cdot (1 - p(a_2) \cdot p(l_{21})) \end{aligned}$$

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 fast. Section 6 will show some experiments comparing the SA approach to the PAS model developed in this section, and demonstrate how probabilities can be computed in practice.

4 Estimating the popularity of a web page

4.1 PageRank

The PageRank algorithm [BL98] 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, since 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 [LG99]. 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 [Mar97], "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. The PAS modeling offers a clean way to take account of these a priori user preferences.

4.2 PAS and personalized "popularity" measure

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 a_i under which d_i corresponds to the user's interest which is denoted as:

$$a_i \to P_i \tag{14}$$

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:

$$P_i \wedge l_{ij} \to P_j \tag{15}$$

Note that the symbolic support that a document belongs to the favorite pages is the same as the support that it will be relevant. In our 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.

5 Finding hubs and authorities

5.1 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 (e.g., "human rights", "MP3"), 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 very 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 [Kle98] is that a good authority is a page which has links from many good hubs, and a good hub is a page which 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 of Kleinberg's HITS algorithm [Kle98] is to consider a root set (e.g., 200 documents), containing the most likely relevant pages found with a search engine in response to a given query. This root set is expanded with all documents which 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:

$$h_p = \sum_{d_p \to d_i} a_i, a_p = \sum_{d_i \to d_p} h_i \tag{16}$$

where $d_p \to d_i$ 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, since 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 [Kle98, CdBD99]. However, we believe there are some weaknesses in this algorithm. One is that a base set has to be chosen for a given topic, from which the hubs and authorities will be selected. Although it is argued in [Kle98] 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 is that the initial ranking of documents is not used as prior evidence, while clearly an initially better ranked document has more chances to be relevant, and certainly more chances to be a good hub or authority.

5.2 A model for computing hub and authority scores

Given a user's request and a document d_i, proposition H_i denotes "document d_i is a good hub" and A_i denotes "document d_i is a good authority". Unlike Kleinberg's algorithm, we consider that there is initial evidence h_i that D_i is a good hub, and a_i that it is a good authority.

$$h_i \to H_i, a_i \to A_i \tag{17}$$

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:

$$H_j \wedge f_{ji} \to A_i \tag{18}$$

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

$$A_i \wedge g_{ij} \to H_j \tag{19}$$

For each hyperlink from d_j to d_i, there will be two rules generated: $(H_j \wedge f_{ji} \to A_i), (A_i \wedge g_{ij} \to H_j)$. 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 $p(a_i)$ 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.

6 Experiments with the model for spreading activation

In this subsection, we present our experiments done on the CACM and Trec'8 web test-collections. The CACM test-collection (3.2 MB) contains 3,204 documents and 50 requests with their associated relevance judgments. The web Track collection, denoted WT2g, contains around 250,000 pages (2.3 GB) extracted from the web and owns 100 requests with their relevance assessments. The retrieval effectiveness is given by the average precision at 11 recall values computed by the TRECEVAL software, the most used evaluation measure in IR.

6.1 Probabilities estimates

The set of symbolic operations is the following:

- Convert each link from d_i to d_j to $D_i \wedge l_{ij} \to D_j$.

Type of link	SA	PAS	PAS vs. baseline	PAS vs. SA
Citing	0.266 (0.3)	0.267	+5.57%	+0.01%
Cited	0.261 (0.4)	0.273	+7.83%	+4.11%

Table 1: Average precision on the CACM test-collection

- For each document D_i, compute the support $sp(D_i, \xi)$.
- Put the support of each document in disjoint form using Heidtmann's algorithm [Hei89]. Convert this disjoint form to a directly usable formula for computing the degree of support $dsp(D_i, \xi)$.

At this point, all the logical operations are done. A learning phase is also necessary to assign probabilities to the assumptions $p(a_i)$ and $p(l_{ij})$. For the first ones, a set of training queries is used to fit a logistic regression to compute the probability of relevance given the page rank. The probabilities of link assumptions can be static or may depend on some similarity value between the link and the search topic. For these experiments, we will only consider static probabilities. They can also be estimated on a set of training queries [Pic98].

When a query is processed, the only operation remaining to compute is the degree of support for each document d_i, $dsp(D_i, \xi)$.

6.2 Experiments

For the CACM collection, links were considered separately in the forward direction (citing) and in the backward direction (cited). For each document, arguments of order three and less were computed, and the symbolic support was put in disjoint form with Heidtmann's algorithm.

The basic retrieval process was done using a classical retrieval system based on the cosine similarity measure and this baseline average precision is 0.253. In Table 1, comparisons were made between PAS and this baseline, and PAS and spreading activation (SA). The results shown for SA are the best ones obtained for a range of values of the parameter λ, which was fixed for all links of a certain type. This best λ value is depicted between parenthesis in Table 1. For the PAS model, arguments of length three or less (at most three links assumptions) were computed.

For the WT2g collection, ten different weighting schemes were used (okapi-npn, ..., bnn-bnn), to produce ten different rankings of documents (for details see [SP01]). Table 2 depicted the retrieval effectiveness using only the best incoming links, the best outgoing links, and both. There are slight but generally not significant improvements over the baseline.

These results demonstrate that PAS can compete with an established technique such as SA. Also, indirect neighbors can be considered without depreciating performance, at least for one test collection. For citing links, there is no average difference in retrieval effectiveness, while for cited links, the difference is nearly significant. However, for SA there is no understanding of the parameters involved, while in the PAS framework, parameters are the probabilities of the links which have a clearer meaning both from a statistical and a logical viewpoint.

7 Discussion

In this paper, we have shown how various approaches using hyperlinks in order to search and organize the web can be modeled using a single theoretical framework. Experiments have validated the approach when links are used to improve an initial ranking. Experiments for computing hub and authorities, and a personalized popularity measure have not been conducted yet because due to the fact that building a good quality test-collection is a very costly and demanding task. But the IR community is tackling the task of building good quality web test-collections, and experiments should be possible in a next future.

Model	Baseline	Best incoming	Best outgoing	Combined
okapi-npn	0.267	0.267 (0.00%)	0.267 (0.00%)	0.267 (0.00%)
lnu-ltc	0.234	0.239 (+2.18%)	0.240 (+2.65%)	0.239 (+2.10%)
atn-ntc	0.257	0.260 (+1.44%)	0.261 (+1.52%)	0.260 (+1.32%)
ntc-ntc	0.138	0.139 (+0.14%)	0.139 (+0.14%)	0.138 (0.00%)
ltc-ltc	0.136	0.138 (+0.88%)	0.138 (+0.88%)	0.138 (+0.88%)
lnc-ltc	0.107	0.108 (+0.65%)	0.109 (+1.49%)	0.107 (+1.31%)
lnc-lnc	0.072	0.074 (+2.35%)	0.073 (+1.38%)	0.073 (+1.52%)
anc-ltc	0.082	0.084 (+2.31%)	0.087 (+5.59%)	0.084 (+2.79%)
nnn-nnn	0.071	0.072 (+0.56%)	0.072 (+0.42%)	0.070 (-0.98%)
bnn-bnn	0.096	0.100 (+4.18%)	0.101 (+5.02%)	0.099 (+3.56%)

Table 2: Average precision on the WT2g test-collection

Because several methods have been suggested to use hyperlinks for different purposes, there is a real need for a formal framework to allow analysis and comparisons between them. We have shown one possible framework here based on propositional logic for translating intuitions into knowledge. With the use of PAS, we have gained some insights on those intuitions: for example, the similarity between the computation of popularity measures and improving an initial ranking have been shown, and we have seen how some heuristic aspects of Kleinberg's algorithm could be described.

References

[Bh98] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 104–111, 1998.

[BL98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the World Wide Web Conf.*, pages 107–117, 1998.

[CdBD99] S. Chakrabati, M. Van der Berg, and B. Dom. Focused crawling: A new approach to topic specific resource discovery. In *Proc. of the World Wide Web Conf.*, pages 545–567, 1999.

[Hei89] K.D. Heidtmann. Smaller sums of disjoint products by subproduct inversion. *IEEE Transactions on Reliability*, 38(3):305–311, 1989.

[HKL99] R. Haenni, J. Kohlas, and N. Lehmann. Probabilistic argumentation systems. Technical Report 99-09, Institute of Informatics, University of Fribourg (Switzerland), 1999.

[Kle98] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the 9th Symposium on Discrete Algorithms*, pages 668–677, 1998.

[LG99] S. Lawrence and C.L. Giles. Accessibility of information on the web. *Nature*, 400(8):107–109, 1999.

[Mar97] M. Marchiori. The quest for correct information on the web: Hyper search engines. In *Proc. of the World Wide Web Conf.*, 1997.

[otCP99] Members of the Clever Project. Hypersearching the web. *Scientific American*, 1999.

[Pic98] J. Picard. Modeling and combining evidence provided by document relationships using probabilistic argumentation systems. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 182–189, 1998.

[Sav94] J. Savoy. A learning scheme for information retrieval in hypertext. *Information Processing & Management*, 30(4):513–533, 1994.

[Sav97] J. Savoy. Ranking schemes in hybrid Boolean systems: A new approach. *Journal of the American Society for Information Science*, 48(3):235–253, 1997.

[SP01] J. Savoy and J. Picard. Retrieval effectiveness on the web. *Information Processing & Management*, 37(4):543–569, 2001.