Information Systems 35 (2010) 530–543

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Modeling the web as a hypergraph to compute page reputation

Klessius Berlt [a,*], Edleno Silva de Moura [a], André Carvalho [a], Marco Cristo [b], Nivio Ziviani [c], Thierson Couto [d]

[a] Department of Computer Science, Federal University of Amazonas, Manaus, Brazil
[b] FUCAPI, Analysis, Research and Tech. Innovation Center, Manaus, Brazil
[c] Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil
[d] Institute of Informatics, Federal University of Goiás, Goiânia, Brazil

ARTICLE INFO

Keywords:
Link analysis
Web search

ABSTRACT

In this work we propose a model to represent the web as a directed hypergraph (instead of a graph), where links connect pairs of disjointed sets of pages. The web hypergraph is derived from the web graph by dividing the set of pages into non-overlapping blocks and using the links between pages of distinct blocks to create hyperarcs. A hyperarc connects a block of pages to a single page, in order to provide more reliable information for link analysis. We use the hypergraph model to create the hypergraph versions of the Pagerank and Indegree algorithms, referred to as HyperPagerank and HyperIndegree, respectively. The hypergraph is derived from the web graph by grouping pages by two different partition criteria: grouping together the pages that belong to the same web host or to the same web domain. We compared the original page-based algorithms with the host-based and domain-based versions of the algorithms, considering a combination of the page reputation, the textual content of the pages and the anchor text. Experimental results using three distinct web collections show that the HyperPagerank and HyperIndegree algorithms may yield better results than the original graph versions of the Pagerank and Indegree algorithms. We also show that the hypergraph versions of the algorithms were slightly less affected by noise links and spamming.

1. Introduction

Modern search engines use algorithms that analyze the web graph to estimate the reputation of each page, and then use this estimate as evidence of the page's relevance when processing queries. The reputation of a page may be interpreted as a measure of the reliability of the page and the importance of its content, according to the opinion of web users. The strategy of computing page reputation, known as link analysis, plays an important role in the quality of the ranking provided by search engines.

Many link analysis strategies have been proposed in the literature in the last decade [7,8,18,16,22]. A common central idea in all the strategies is that a link from a page to another may represent a vote for the reputation of the destination page. The first method proposed to exploit this source of information was Indegree, which uses the number of links to a page as an estimation of its reputation [7]. This first idea was then followed by more sophisticated strategies, with the Pagerank [8] being one of the most known and successful of them. All these approaches adopt a graph model for the web, where vertices represent pages and edges represent the links between pages.

One of the main problems in link analysis strategies is to determine whether a link to a page can be considered as

[*] Corresponding author. Tel.: +55 9284137845.
E-mail addresses: klessius@gmail.com (K. Berlt), edleno@dcc.ufam.edu.br (E.S. de Moura), andre@ufam.edu.br (A. Carvalho), marco.cristo@fucapi.br (M. Cristo), nivio@dcc.ufmg.br (N. Ziviani), thierson@inf.ufg.br (T. Couto).

0306-4379/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2009.02.005

a vote for its quality or not. Some links, such as navigational purpose links or spam links, can lead link analysis methods to wrong conclusions about the page reputation when these links are considered as votes. Another example of links that may mislead link analysis algorithms is the occurrence of redundant link information that may be taken as distinct votes. For instance, if several pages of a site link to a page *p*, then all these pages will contribute to the reputation of *p*. We argue that this influence may be considered as negative, since it does not necessarily reflect the opinion of the whole web community. Links from a single site are not likely to represent independent opinions about the reputation of a page. A possible solution to this problem is to produce models that are able to reduce the impact of links originating from the same site, giving more importance to the diversity of links received by a page.

In this work we propose a representation of the web as a hypergraph, instead of a regular graph, producing a model where the connections are less redundant, in the sense that they are more likely to be independent from each other. The proposal aims at producing better page reputation values. We show how to adapt the Pagerank method [8] and the Indegree method [7] to compute page reputation in this new model. We call the hypergraph versions of the two methods HyperPagerank and HyperIndegree, respectively. In our model, the web hypergraph is derived from the web graph by partitioning the set of pages, that is, by dividing it into non-overlapping blocks. The links between pages found in the original web graph are used to define the set of hyperarcs in our hypergraph. Given a block of pages $\mathscr{B}$ and a page *p*, there is an hyperarc from $\mathscr{B}$ to *p* if and only if there is one or more web links from pages of $\mathscr{B}$ to *p*.

Using this abstraction to represent the web it is possible to derive a family of new methods for computing page reputation by adapting traditional methods to use the hypergraph representation. The key difference between our approach and the traditional web graph representation is that our model aims to represent connections that are more likely to be independent of each other than the links in the web graph. The hypergraph model allows the control of the influence of the individual page connections on a vote. The more fine-grained the page blocks are, the greater is the influence of the link on their votes. This is a key point because it gives the model flexibility to deal with the differences in the quality of the links.

The criterion used to group pages in a block can vary according to the final goal of the link analysis method. We argue that, at the same time, the chosen granularity should be coarse-grained enough to allow independent votes and fine-grained enough to maintain a sufficient number of hyperarcs to enable the computation of the page reputation. When searching in a collection with a large number of web sites, the optimal point in this trade-off is not likely to take into account all the links used in the traditional representation of links between pages.

We present experimental results varying the partition criteria to show the impact of different levels of partition granularity. We used two web collections of documents. We combined page reputation, textual content of pages and anchor text information available in the collections used in the experiments. In one of the collections the hypergraph versions of the link analysis methods considered here produced a significant improvement in the ranking quality for navigational queries (queries in which the user looks for a specific web site), and maintained a similar ranking quality to that of the graph-based versions for informational queries (queries in which the user looks for some specific information).

In the other web collection the information about navigational and informational queries were not available and the hypergraph versions of the algorithms did not yield significant changes in the results for both navigational and informational queries. We investigated the possible reasons for such differences and concluded that they were due to the fact that the second collection has less connectivity information than the first one. These results show that the hypergraph model is an interesting alternative to consider in search applications that adopt page reputation algorithms.

This work is organized as follows. Section 2 discusses the related work. Section 3 describes the hypergraph model. Section 4 presents the implementation of the hypergraph versions of Pagerank and Indegree. Section 5 presents the experimental setup. Section 6 presents experimental results comparing the hypergraph versions of Pagerank and Indegree to their original graph versions. Section 7 discusses the impact of spam on both versions of the algorithms. Finally, Section 8 presents the conclusions and discuss future research directions.

2. Related work

One of the first efforts to analyze the link structure of the web and use it as a source of evidence in web search engines was the Indegree method [7]. The method uses the number of incoming links of a page as a heuristic for determining the importance of each web page. The intuition behind is that pages with more incoming links have more visibility, and thus may also have a high reputation in terms of quality.

Pagerank [8] is one of the most successful link analysis method. It computes a web page reputation score as the probability of a random surfer reaching that page. As opposed to the Indegree method, an important characteristic of the Pagerank method is that a web page may have a high reputation without having a high number of links pointing to it, since links from pages that have high Pagerank scores have high influence on the final Pagerank scores of other pages. This characteristic can be seen as an advantage, since pages with high Pagerank scores, which tend to be high quality pages, have more influence on the final results than pages with low Pagerank scores. However, this characteristic of Pagerank might create a bias, since a page can receive a high Pagerank score even though it is pointed by a small number of pages [10]. A thorough study of the properties of Pagerank is presented in [17].

The problem of a few pages having a strong influence on the final scores of other pages also appears in other important and popular link analysis methods, such as the HITS method proposed by Kleinberg [16], its variant proposed by Bharat and Henzinger [5] and the SALSA method [18], which was inspired by HITS and Pagerank. The work of Borodin et al. [6] also analyzes the properties and performance of some of the most popular link analysis algorithms.

Bharat and Henzinger [5] try to solve the problem of a few pages having influence on other pages by giving weights to the edges according to the host of the source page. If there are k links from pages in a host H to a page p, each link will have a weight of $1/k$, assuring that pages from the same host will have a limited influence on the authority and hub values of the pages. They evaluated their ideas in a context of local link analysis in which the web graph is formed on the fly from the answer set of each query. As we were interested in the study of global link analysis in which the whole web graph is considered we did not include their results in our study.

Amento et al. [1] present experiments for a site-level version of Indegree, where all the links to a web site are computed as links to its root page. The idea of computing the reputation by considering high-level entities, such as sites or domains, is in fact explored in many previous works [7,13,4,2]. However, as far as we know, no previous work has presented a model that allows the representation of both pages and high-level entities together as we do here. One of the advantages of this representation is to allow an easy adaptation of previously proposed link analysis methods to deal with high-level entities when computing page reputation. Also, our model can be easily extended to represent different page partitions. For instance, we could partition the collection such that all the pages belonging to a link farm would be treated as a unique entity, diminishing their influence.

3. The hypergraph model

Our model uses a directed hypergraph to represent the web. A directed hypergraph $\mathscr{H} = (\mathscr{V}, \mathscr{E})$ consists of a set of vertices $\mathscr{V}$ and a set of hyperarcs $\mathscr{E}$, where $\mathscr{E} \subseteq 2^{\mathscr{V}} \times 2^{\mathscr{V}}$. Since we intend to calculate the importance of individual pages, we redefine $\mathscr{E}$ in a more restrictive way, that is, $\mathscr{E} \subseteq 2^{\mathscr{V}} \times \mathscr{V}$. Thus the hyperarcs always point to single vertices and each hyperarc $\varepsilon = (G, v) \in \mathscr{E}, G \subseteq 2^{\mathscr{V}}$, we have that $v \notin G$. We have also considered the representation of the hypergraph without the last restriction ($v \notin G$), which is equivalent to allowing internal hyperlinks, but the final results were similar to the version with this restriction.

To model the web we consider each page as a vertex of the graph and partition the set of pages into non-overlapping page blocks, where the pages are grouped according to an affinity criterion. In our model a partition block $\mathscr{B}$ in the hypergraph points to a page v through a hyperarc $\varepsilon = (\mathscr{B}, v)$ if and only if there is at least one page of $\mathscr{B}$ that has a web link to the page v and $v \notin \mathscr{B}$. An important difference between this model and the traditional web graph model is that the partition criterion determines the granularity of the hyperarcs.

In our model the partition of the web graph is based on the hierarchy derived from the URLs of pages. The advantage of partitioning the collection based only on the URL hierarchy is the low cost of this procedure. We considered the following three alternative partition criteria based on the URL hierarchy:

- *Page-based partition*: A block is composed of a single web page. This criterion is the traditional graph representation of the web.
- *Domain-based partition*: All pages of a block belong to the same web domain.
- *Host-based partition*: All pages of a block belong to the same web host.

We are able to simulate the traditional web representation by using a page-based partition, where each page is treated individually. By doing so, our system is able to simulate traditional link analysis methods, providing appropriate comparison baselines for our experiments. Thus, this criterion is represented in the experiments by the graph versions of the methods studied.

The adopted domain-based and host-based partitions group pages that are probably created by the same author or by related authors. This possibility was mentioned in the literature [7,13] as an option to compute the Indegree, but no actual evaluation of its impact on the ranking of web search engines was performed. Further, the chance of two hosts or two domains being created by the same author or by related authors is smaller than that for pages. As a result, we expect sets of partition blocks that are highly reliable, since the page reputation computed by considering hyperarcs coming from a domain or a host is proportional to the *diversity* of partition blocks that point to the page rather than to the number of links to the page, as is the case in the traditional representation of the web graph.

These three partition criteria are implemented using the URL of the pages. The page-based partition is directly determined by the distinct URLs of the collection. For the host-based and domain-based partitions, we use *domain names* and *host names*. The definition of host and domain names here is based on a string matching process applied to the URL of each page.

To obtain a host name, the URL is first processed to remove the starting prefixes “http://” and “www.”. Next, the host name is defined as the string starting at the beginning of the resulting URL and finishing at the position before the first slash. In the URL “http://dir.yahoo.com/”, for example, the host name is “dir.yahoo.com”. To obtain a domain name we first divide the host name into parts, according to the dots found on it. Then, we obtain the country id, such as “.fr” and “.br”. For certain pages, the country id may be empty. We then obtain the server category, such as “.com” and “.edu”, which also may be empty. At the end, we take the last part

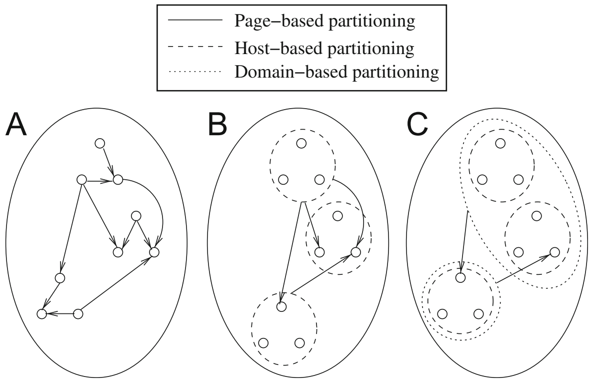

Fig. 1. The hypergraph model considering (A) pages as the partition criterion, (B) hosts as the partition criterion and (C) domains as the partition criterion.

of the server name, which is neither a category nor a country id, to be the domain core. Finally, the domain name is defined as the concatenation of the domain core, the category, and the country id. For instance, in the URL "http://dir.yahoo.com", the domain name is "yahoo.com". In the URL "http://www.uol.com.br/esportes/~index.html", the domain name is "uol.com.br".

In the hypergraph model, every URL is parsed as above and three partitions are considered, as illustrated in Fig. 1. Fig. 1(A) represents the page-based partition (simulating the traditional web graph representation). Fig. 1(B) represents the host-based partition, which groups pages with the same host name in one block. Fig. 1(C) represents the domain-based partition, which groups together pages with the same domain name, each domain consisting of a set of one or more hosts. The number of hyperarcs decreases as the partition criterion includes a larger number of elements. The selection of each level of granularity creates a trade-off between the number of hyperarcs (and thus the amount of information about each page) and the qualitative information provided by each hyperarc.

We also considered the use of another partitioning strategy that uses a clustering algorithm to group the pages. We investigate the use of the minimum spanning trees-based clustering algorithm (MST) [23] to produce clusters. The similarity measure adopted in our experiments was the intersection between the links from each group. Considering that the results obtained were not superior to the ones obtained by the URL match partition criteria and the computational costs were prohibitive, we decided not to include these experiments in this paper. However, the study of other similarity measures and other clustering algorithms could still yield improvements in further studies.

4. Implementation of the algorithms using the hypergraph model

In this section we discuss the implementation of the Indegree and Pagerank algorithms using the hypergraph model. The Indegree method was chosen because it is a simple and effective link analysis method. The Pagerank method was chosen because it is considered a successful link analysis method. It is also usually adopted as a baseline for link analysis in the literature.

4.1. Indegree and Pagerank methods

The Indegree method consists of counting, for each page p, the number of incoming links to p. While this is quite a naive method that is susceptible to noise and spam, it is useful to provide a further example of how an appropriate choice of the partition criterion in the hypergraph model can improve the quality of link analysis methods.

The Pagerank method computes the reputation of a page as the probability of a random surfer visiting that page. Given a page p, the Pagerank formula is

$$PR(p) = (1-c) \times \sum_{q \in I(p)} \frac{PR(q)}{\|O(q)\|} + \frac{c}{\|\mathscr{V}\|} \tag{1}$$

where c is the dampening factor, $I(p)$ is the set of pages that point to page p, $\|O(q)\|$ is the number of pages pointed by q and $\|\mathscr{V}\|$ is the number of pages in the collection.

Since the hypergraph approaches using host and domain names naturally disregard internal links, we implemented variations of Pagerank and Indegree that do not consider such links in the web graph. Thus, we implemented three distinct versions of the Pagerank and Indegree methods:

(1) The original page-based graph version considering all links, referred to as Pagerank and Indegree.
(2) A host-based version considering only links between pages in distinct hosts, referred to as PRHost and IndHost.
(3) A domain-based version considering only links between pages in distinct domains, referred to as PRDom and IndDom.

These variations are useful to check whether the improvements achieved using the hypergraph model are due to the removal of internal links or not.

4.2. Computing reputation in the hypergraph model

An important step is to adjust the link analysis methods from the traditional graph representation to the hypergraph model. We show here how to do it for the two algorithms:

(1) Pagerank, which leads to HyperPagerank, used with two different granularities: Using the domain as the partition criterion (referred as HyPRDom), and using the host as the partition criterion (referred as HyPRHost).
(2) Indegree, which leads to HyperIndegree, used with two different granularities: Using the domain as the partition criterion (referred as HyIndDom) and using

the host as the partition criterion (referred as HyIndHost).

Other variations of these methods can also be derived for the hypergraph model. In Section 6 we present experiments to evaluate the performance of these two methods when compared to their original versions.

To compute the HyperPagerank we need a way to compute the Pagerank values of each partition block in order to compute the Pagerank of each page. A simple and effective strategy is to consider the reputation of a block as the sum of the reputation of all pages in that block. We compute each Pagerank iteration in two steps:

(1) The reputation of each block of pages $GR(\mathscr{B})$ is computed as the sum of the reputation of each page p that belongs to it:

$$GR(\mathscr{B}) = \sum_{p \in \mathscr{B}} PR(p) \tag{2}$$

where $PR(p)$ is the current Pagerank value of page p.

(2) The Pagerank value of each page depends on its representation in the hypergraph. Pages with no incoming hyperarcs have value 0, meaning that they have no reputation. For each page p with incoming hyperarcs, we give an initial value $1/\|\mathscr{V}\|$, where $\|\mathscr{V}\|$ is the number of pages with incoming hyperarcs in the collection, and compute the reputation $PR(p)$ as

$$PR(p) = (1-c) \times \sum_{\mathscr{B} \in I(p)} \frac{GR(\mathscr{B})}{\|O(\mathscr{B})\|} + \frac{c}{\|\mathscr{V}\|} \tag{3}$$

where c is the dampening factor, $I(p)$ is the set of page blocks that point to page p, and $\|O(\mathscr{B})\|$ is the number of pages pointed by block $\mathscr{B}$.

These two steps are repeated until the values converge. Note that, as in the original Pagerank, the convergence is assured. This hypergraph version of Pagerank may also be interpreted as computing the Pagerank of a site with the propagation of values to the pages of each site being performed according to the hyperlinks received by each page.

It is also important to stress that, while in Pagerank every page has a minimum Pagerank value, in the HyperPagerank pages that do not receive at least one hyperarc would have an HyperPagerank value of 0. This was done because if all pages received a minimum value, this would bias the HyperPagerank method towards giving large authority values to blocks with a lot of pages, regardless of the quantity of hyperarcs their internal pages receive.

The computation of our version of HyperIndegree is straightforward and consists of counting, for each page p, the number of hyperarcs that reach it. Thus, the HyperIndegree value of p is computed as

$$HI(p) = \sum_{\mathscr{B} \in I(p)} 1 \tag{4}$$

where $I(p)$ is defined as in Eq. (3).

Table 1
Statistics about the WBR03 collection.

No. of pages	12,020,513	No. of hosts	1,001,070
No. of domains	141,284	No. of links	130,717,004
No. of hyperarcs (host)	32,414,004	No. of hyperarcs (domain)	1,906,879
Average plain text size	5 Kb		

5. Experimental setup

We conducted experiments to assess the impact of using the hypergraph model to compute page reputation in search tasks. We used two distinct web collections for the ranking experiments: WBR03 and WT10g.

5.1. The WBR03 collection

The WBR03 collection is a real search engine database of the search engine TodoBR,[1] composed of 12,020,513 web pages collected from the Brazilian Web in 2003. As depicted in Table 1, the WBR03 collection has 130,717,004 valid links between its pages and the average size of plain text of each document is 5 Kb. This number of valid links indicates that WBR03 has a highly connected set of pages, providing rich information for link analysis methods. It represents a considerably connected snapshot of the Brazilian Web community, which is probably as diverse in content and link structure as the entire Web. Thus, we believe it makes a realistic testbed for our experiments. Table 1 presents more information about the WBR03 collection.

In the experiments we used queries extracted from a log of 3 million queries submitted to TodoBR in order to evaluate the impact of our methods in practical situations. We divided the query set into two main groups:

(1) *Navigational queries*: Where the user is searching for a specific web site.
(2) *Informational queries*: Where the user is searching for information on a given topic.

Each group was divided into popular queries and randomly selected queries. Thus, we performed experiments with four distinct query sets. All these sets of queries were evaluated by 15 people, all of them familiar with the Brazilian Web, in order to ensure more reliability to our experiments.

The set of popular navigational queries was composed of the 60 most popular navigational queries found in the log (a sample set of 15 of the queries used is given in Table 16 in the Appendix). The set of randomly selected navigational queries was composed of 60 queries randomly selected from the log (a sample set of 15 of the queries used is given in Table 17 in the Appendix). For all

[1] TodoBR is a trademark of Akwan Information Technologies, which was acquired by Google in July 2005.

navigational queries, the results were evaluated using the metric MRR (mean reciprocal ranking), which is the most common metric for evaluating the quality of results in navigational queries, being also the metric adopted for navigational queries on the TREC Conference.[2]

The MRR is calculated as

$$MRR(QS) = \frac{\sum_{\forall q_i \in QS} \frac{1}{PosCorrectAnswer(q_i)}}{|QS|} \qquad (5)$$

where *QS* is a query set submitted to a system and *PosCorrectAnswer* is the position where the first correct answer for the query q_i was found. The MRR value of a system is a real number between 0 and 1, where 1 is the best MRR value possible, obtained only if all the correct answers are placed in the first position in the ranking of the answers returned by the system.

The set of popular informational queries contained the 50 most popular queries found in the log (a sample set with 15 of the queries used is detailed in Table 18 in the Appendix). The set of randomly selected informational queries was composed of 50 queries (a sample set with 15 of the queries used is detailed in Table 17 in the Appendix). We evaluated informational queries using the same pooling method used within the TREC web collection [15]. We thus constructed query pools containing the top 20 answers for each query and method. Then, we assessed our output in terms of various precision-based metrics. For each method, we evaluated the mean average precision (MAP) and the precision at the first 10 positions of the resulted ranking (P@10).

We processed both the navigational and the informational queries according to the user specifications, as extracted from the log: phrases, Boolean conjunctive or Boolean disjunctive.

5.2. *The WT10g collection*

The second collection used for the search experiments is WT10g, a collection adopted in the TREC 2001 web track [3]. As can be seen in Table 2, the WT10g collection contains 1,692,096 documents extracted from the whole web. The average size of plain text after parsing the documents is about 4.4 Kb. The number of links between pages in the collection is 2,530,920. We observe that the number of links between pages in the WT10g collection (roughly 1.5 times the number of pages) is significantly smaller than this number in the TodoBR collection (roughly 11 times the number of pages). This difference may affect the choice of link analysis methods on each collection, and it is important to conduct experiments in two distinct scenarios.

Other collection characteristics that may affect the choice of link analysis methods include the relations between number of pages, hosts and domains on each collection. The difference between the number of domains and hosts in the WT10g collection is only 10%, which is small when compared to this difference in the WBR03 collection. This small difference may affect the behavior of the hypergraph algorithms in the WT10g collection, since it reduces the differences between the domain-based and host-based partition criteria. For the experiments with the WT10g collection, we adopted the first 145 homepage finding queries, which are all navigational queries (a sample set with 15 of the queries used is detailed in Table 20 in the Appendix). Informational query results are not presented in the experiments for this collection, since the quality of the results, as in WBR03, was not affected by the link analysis methods adopted in this collection.

Table 2
Statistics about the WT10g collection.

No. of pages	1,692,096	No. of hosts	11,671
No. of domains	10,113	No. of links	2,530,920
No. of hyperarcs (host)	1045	No. of hyperarcs (domain)	952
Average plain text size	4.4 Kb		

The experiments with the WT10g collection are useful to study the behavior of the methods on a collection with a small number of links. Note that this is quite a difficult scenario for the hypergraph model, since the total number of hyperarcs is even more reduced than the total number of links.

6. Experimental results

We tested the methods in two scenarios for both collections:

(1) Only the page reputation algorithms were applied for computing the ranking of the pages that contained the terms of the query, thus avoiding interferences from evidence combination in the comparison results. In this case the results will be referred to as *no combination*.
(2) We combined the link analysis method with the result of the vector space model [19] over the textual content of the web pages and over the anchor text information, where each page is represented by the concatenation of all the anchor text found in links that point to it.

We experimented the second scenario with two different combination approaches:

(1) A Bayesian belief network framework, as described in [9], referred to as *BNC*.
(2) A brute force training-based combination method described in [12], referred to as *BFC*.

The BFC scenario is useful to provide a better idea about the impact of the methods in a practical situation. We also tested the combination without link analysis in order to assert the impact of the link analysis methods in the final

[2] http://trec.nist.gov/

ranking. In all the experiments, we used a *t*-test to evaluate the significance of the results, as suggested in [20]. The significance level adopted was 95%, and results with less than this threshold were considered as non-conclusive.

To provide information about the individual impact of each method on the final ranking quality we also present the results of the methods with no combination at all (where all the pages containing the words of the query are ranked according to their reputation scores given by each link analysis method), which is referred to as *no combination*. However, it is important to note that page reputation is a query-independent information, and thus a ranking with no combination does not make much sense. Results with *no combination* are provided only for comparison of the relative impact of each method in each query set.

6.1. Experiments with the WBR03 collection

In this section we used the WBR03 collection with four distinct types of query: popular navigational queries, randomly selected navigational queries, popular informational queries, and randomly selected informational queries. Then, we perform experiments to evaluate the impact of a noise removal method on the results obtained with this collection.

Tables 3 and 4 present the MRR results when processing the set of popular navigational queries with the Pagerank versions and the Indegree versions, respectively. The results indicate that the hypergraph versions of the methods in both cases are superior to the graph versions for this set of queries. We applied a *t*-test and found that all the differences between the hypergraph versions of Pagerank and Indegree and their corresponding graph versions are significant.

In the case of most popular navigational queries, the results were improved when using the domain-based and host-based methods. An example of change in results can be seen when processing the query "BOL",[3] where the first two results provided by Indegree are blog pages pointed by many other blog pages. The right answer appears in the third position of Indegree. The HyIndDom method was not affected in this case because, as expected, the number of distinct domains pointing to the two mentioned blog pages is far smaller than the number of distinct domains pointing to the BOL home page, which is the right answer and is pointed by a more diverse set of people.

The differences in results between PRDom and HyPRDom and between IndDom and HyIndDom are useful to show that the improvements are not only due to the removal of internal links in the hypergraph, but a consequence of the better representation of connections provided by the hypergraph model. For instance, PRDom was the best Pagerank implementation when modeling the web as a graph for navigational queries, which indicates that the removal of internal links has a positive effect on navigational queries in the WBR03 collection. However, for popular navigational queries, the HyPRDom method achieved a gain of 30.48% compared to PRDom when using only link analysis, 23.60% when using BNC to combine other pieces of evidence, and 25.42% when using BFC. As in HyIndDom, the results indicate that its performance is not only due to the removal of internal links.

Table 3
Pagerank versions: mean reciprocal rank (MRR) values for popular navigational queries in the collection WBR03, modeling the web as a graph (Pagerank, PRHost and PRDom) and as a hypergraph (HyPRHost and HyPRDom).

Pagerank versions (popular navigational queries)			
Method	No combination	BNC	BFC
Pagerank	0.2834	0.4610	0.4553
PRHost	0.3987	0.5036	0.5794
PRDom	0.4888	0.5785	0.6361
HyPRHost	0.5535	0.5819	0.6174
HyPRDom	0.6378	0.7150	0.7978

Table 4
Indegree versions: mean reciprocal rank (MRR) values for popular navigational queries in the collection WBR03, modeling the web as a graph (Indegree, IndHost and IndDom) and as a hypergraph (HyIndHost and HyIndDom).

Indegree versions (popular navigational queries)			
Method	No combination	BNC	BFC
Indegree	0.5109	0.5890	0.6726
IndHost	0.5413	0.6101	0.6565
IndDom	0.6510	0.7122	0.7617
HyIndHost	0.5964	0.6241	0.7166
HyIndDom	0.7273	0.7706	0.8547

Tables 3 and 4 also show that the results obtained when considering domains as the partition criterion were superior to the results obtained when using hosts. This same conclusion is also obtained when testing the other query types in WBR03. We examined the partitions created when using hosts as the partition criterion in order to investigate the reasons for its poor performance. We found out that it creates many connected partition elements due to the replication of hosts with different names, or due to strongly related hosts, such as different hosts from a same web portal. For instance, the hosts "http://esportes.uol.com.br" and "http://games.uol.com.br" are from the same portal, and thus have hyperarcs connecting them to each other. These cases are quite common in the web and create hyperarcs that are not likely to be considered as votes for quality. As a consequence, they reduce the quality of the results obtained when using hosts as the partition criterion.

Note that in all cases, the differences between the methods are attenuated when the ranking function uses other evidence; however, the hypergraph versions of the link analysis strategies still result in improvements compared to their original versions over the web graph. The results for the popular navigational query set indicate

[3] BOL(http://www.bol.uol.com.br/) is one of the largest Brazilian web sites.

Table 5
Pagerank versions: mean reciprocal rank (MRR) values for randomly selected navigational queries in the collection WBR03, modeling the web as a graph (Pagerank, PRHost and PRDom) and as a hypergraph (HyPRHost and HyPRDom).

Pagerank versions (random navigational queries)			
Method	No combination	BNC	BFC
Pagerank	0.2823	0.5729	0.5280
PRHost	0.3599	0.5614	0.6442
PRDom	0.4642	0.6216	0.6967
HyPRHost	0.4899	0.6144	0.6834
HyPRDom	0.5562	0.6889	0.7856

Table 6
Indegree versions: mean reciprocal rank (MRR) values for randomly selected navigational queries in the collection WBR03, modeling the web as a graph (Indegree, IndHost and IndDom) and as a hypergraph (HyIndHost and HyIndDom).

Indegree versions (random navigational queries)			
Method	No combination	BNC	BFC
Indegree	0.4353	0.6258	0.7468
IndHost	0.4540	0.5832	0.7470
IndDom	0.5256	0.6587	0.7916
HyIndHost	0.5236	0.5863	0.7841
HyIndDom	0.6343	0.6784	0.8391

Table 7
Cross-validation for popular navigational queries.

Popular queries						
Method/fold	1	2	3	4	5	Avg.
Indegree	0.6726	0.6205	0.5169	0.6066	0.5236	0.5880
IndHost	0.6565	0.5853	0.5069	0.6152	0.6095	0.5947
IndDom	0.7617	0.7572	0.7241	0.7655	0.6412	0.7299
HyIndHost	0.7166	0.6516	0.5622	0.6649	0.6388	0.6468
HyIndDom	**0.8547**	**0.8548**	**0.7964**	**0.8115**	**0.8391**	**0.8313**
Pagerank	0.4553	0.5041	0.5094	0.5322	0.4848	0.4972
PRHost	0.5794	0.5428	0.5169	0.5783	0.5121	0.5459
PRDom	0.6361	0.6251	0.6507	0.6748	0.6208	0.6415
HyPRHost	0.6174	0.6325	0.5373	0.6332	0.5595	0.5960
HyPRDom	**0.7978**	**0.7861**	**0.7888**	**0.7802**	**0.6271**	**0.7560**

The best MRR value for each training set is in bold.

Table 8
Cross-validation for randomly selected navigational queries.

Randomly selected queries						
Method/fold	1	2	3	4	5	Avg.
Indegree	0.7468	0.6501	0.7535	0.6277	0.6434	0.6843
IndHost	0.7470	0.6131	0.7315	0.6160	0.6333	0.6682
IndDom	0.7916	0.7964	0.8273	0.6953	**0.6914**	0.7604
HyIndHost	0.7841	0.6735	0.7185	0.6446	0.5539	0.6749
HyIndDom	**0.8391**	**0.8309**	**0.8613**	**0.7963**	0.6648	**0.7984**
Pagerank	0.5280	0.4824	0.4827	0.5701	0.5515	0.5229
PRHost	0.6442	0.4774	0.6122	0.5809	0.5538	0.5737
PRDom	0.6967	0.6614	0.6215	0.6557	0.6800	0.6631
HyPRHost	0.6834	0.6864	0.7242	0.6770	0.6089	0.6760
HyPRDom	**0.7856**	**0.7534**	**0.8238**	**0.7539**	**0.7491**	**0.7731**

The best MRR value for each training set is in bold.

Table 9
MAP and P@10 values for the popular informational queries in the collection WBR03, modeling the web as a graph (Pagerank, PRHost and PRDom) and as a hypergraph (HyPRHost and HyPRDom).

Pagerank versions (popular informational queries)						
Method	No combination		BNC		BFC	
	MAP	P@10	MAP	P@10	MAP	P@10
Pagerank	0.064	0.334	0.105	0.456	0.428	0.643
PRHost	0.058	0.300	0.095	0.412	0.489	0.757
PRDom	0.053	0.298	0.098	0.422	0.487	0.757
HyPRHost	0.067	0.370	0.099	0.434	0.481	0.753
HyPRDom	0.057	0.312	0.093	0.410	0.498	0.777

that the proposed model is especially useful for this type of query. Thus, this information could be used to improve the results in a search system that automatically classifies the queries submitted to it according to their popularity.

Tables 5 and 6 present the results obtained when testing the methods with the set of randomly selected navigational queries. The results with randomly selected queries measures the expected gain in cases where the search engine uses a single ranking function for all navigational queries. As can be seem in Table 5, the hypergraph versions of Pagerank, HyPRHost and HyPRDom, still achieve better results when compared to the graph versions of Pagerank. When comparing the hypergraph and graph Indegree versions presented in Table 6, the results are quite close, with a slight advantage for the hypergraph versions when using any combination method. In the case with no combination, the best results were obtained by the hypergraph versions of Indegree.

We also performed a cross-validation test to verify the impact of the choice of the set of queries used in the training phase of the BFC combination. The pool of 75 queries was divided into five folds, each one containing 15 queries, so when one fold was chosen to be the training set, the four others comprised the testing set. The results are shown in Tables 7 and 8.

Tables 7 and 8 show that, between the Pagerank versions, HyPRDom is the one with the best overall results, yielding an average performance 17.85% better than PRDom on popular queries and 16.59% on randomly selected queries. Also, HyIndDom produced the better MRR results between the Indegree algorithm versions, with average gains of 13.89% over IndDom on popular queries and 5% on randomly selected queries.

Tables 9 and 10 depict the results obtained for the informational queries. *t*-Test results obtained from comparisons between the graph and hypergraph versions of the methods indicate that there is no significant difference in all the comparative results. The range of the results is tighter than for the navigational queries, because in this

Table 10
Indegree versions: MAP and P@10 values for the popular informational queries in the collection WBR03, modeling the web as a graph (Indegree, IndHost and IndDom) and as a hypergraph (HyIndHost and HyIndDom).

Indegree versions (popular informational queries)						
Method	No combination		BNC		BFC	
	MAP	P@10	MAP	P@10	MAP	P@10
Indegree	0.058	0.316	0.108	0.486	0.488	0.763
IndHost	0.053	0.302	0.087	0.394	0.473	0.737
IndDom	0.056	0.306	0.081	0.368	0.487	0.763
HyIndHost	0.056	0.318	0.099	0.452	0.473	0.736
HyIndDom	0.066	0.364	0.105	0.428	0.495	0.780

Table 11
Indegree versions: mean reciprocal rank (MRR) values for navigational queries in the collection WBR03 when using the experimented methods after removing noisy links.

Indegree versions			
Method	No combination	BNC	BFC
Popular queries			
IndDom	0.6940	0.7256	0.7617
HyIndDom	**0.7357**	**0.7592**	**0.8547**
Randomly selected queries			
IndDom	0.5738	0.6277	0.7096
HyIndDom	**0.5891**	**0.6446**	**0.7591**

Table 12
Pagerank versions: mean reciprocal rank (MRR) values for navigational queries in the collection WBR03 when using the experimented methods after removing noisy links.

Pagerank versions			
Method	No combination	BNC	BFC
Popular queries			
PRDom	0.5556	0.6242	0.6361
HyPRDom	**0.6639**	**0.7178**	**0.7978**
Randomly selected queries			
PRDom	0.4669	0.6366	0.6626
HyPRDom	**0.5181**	**0.6407**	**0.7475**

Table 13
Pagerank versions: mean reciprocal rank (MRR) values for navigational queries in the collection WT10g, modeling the web as a graph (Pagerank, PRHost and PRDom) and as a hypergraph (HyPRHost and HyPRDom).

Pagerank versions			
Method	No combination	BNC	BFC
Pagerank	0.0840	0.2547	0.3224
PRHost	0.0762	0.2883	0.3196
PRDom	0.0762	0.2883	0.3196
HyPRHost	0.0218	0.2874	0.3196
HyPRDom	0.0218	0.2868	0.3188

case users are more interested in the content of the answer pages, which makes the page content more important than in the navigational queries. Thus, a lower variation in the quality of results is expected when changing only the link analysis strategies for this query type.

Another important detail is that the training performed by BFC resulted in a function that gives a low weight to the page reputation. We implemented a combination without using link analysis, and the quality of results was similar to the ones obtained with BFC including link analysis, meaning that the page reputation has a low impact on this type of query. We also performed experiments with the randomly selected query set of informational queries, and all the conclusions were equivalent to the ones obtained for popular informational queries. Thus we decided not to show the tables for the non-popular informational queries to avoid too much repetition.

A final experiment with the WBR03 collection was performed to evaluate the impact of noise links on the results obtained by each method. The results obtained in the experiments above consider the WBR03 collection without any pre-processing. One could argue that if a noise removal algorithm had been applied to the collection, then the advantages of the hypergraph model would have been reduced. To check this possibility, we have repeated the experiments with the best methods (IndDom, HyIndDom, PRDom and HyPRDom) with navigational queries on the WBR03 collection, but now applying noise link removal techniques proposed in [10], where it is shown that such spam-removing techniques can improve the effectiveness of link analysis algorithms on navigational queries in the WBR03 collection.

The results for both Indegree and Pagerank versions are shown in Tables 11 and 12, respectively. As can be seen, the hypergraph methods still obtain better results for the WBR03 collection, even when noise links are removed. Thus, the hypergraph model-based methods yield better results on navigational queries of WBR03 in both scenarios, with and without applying noise removal algorithms.

In summary, the experiments with WBR03 indicate that the use of the hypergraph model can be considered as a very good alternative for link analysis, since it is the best option for navigational queries and has equal performance to the other methods in informational queries.

6.2. Experiments with the WT10g collection

To evaluate the performance of the proposed algorithms on another scenario, we also performed experiments with a small collection, the WT10g collection. We used the 145 homepage-finding queries of WT10g, which are all navigational queries. Note that this collection has no query log; thus, there is no way of dividing the queries according to their popularity. Tables 13 and 14 show that all the methods yield unsatisfactory results when used as the unique source of information. The hypergraph algorithms seems to produce worse results when compared to

Table 14
Indegree versions: mean reciprocal rank (MRR) values for navigational queries in the collection WT10g, modeling the web as a graph (Indegree, IndHost and IndDom) and as a hypergraph (HyIndHost and HyIndDom).

Indegree versions			
Method	No combination	BNC	BFC
Indegree	0.0616	0.1901	0.3299
IndHost	0.0776	0.2698	0.3147
IndDom	0.0776	0.2698	0.3147
HyIndHost	0.0776	0.2698	0.3188
HyIndDom	0.0776	0.2698	0.3188

their original graph versions. This can be explained by the fact that the hypergraph lost some information when generating the hyperarcs. However, this difference is not present when the methods are combined with the text and the anchor text of the page. The results in BNC and BFC are very similar for all the methods because this collection provides little link information (only 1.5 links/page on average).

For the WT10g, the experiments indicate that the use of the hypergraph model did not result in a significant change in the ranking quality. This conclusion was expected, since there is not much link information in this collection.

7. The influence of spam in the methods

In this section we study the behavior of the best configuration of hypergraphs obtained in the experiments considering spam pages. We then compare this behavior with that achieved by the methods that use the traditional graph model. In Section 7.1 we present the WEBSPAM-UK2006 collection used in the experiments. In Section 7.2 we present the experimental results involving spam pages.

7.1. The WEBSPAM-UK2006 collection

The WEBSPAM-UK2006 Database[4] is a large collection of pages, links and annotations about spam/non-spam hosts, labeled by a group of volunteers. The collection contains 77,741,046 pages in 11,402 hosts and 7650 domains, all belonging to the .UK country domain. It has almost 3 billion links between pages, about 11 million links when considering only inter-host links and about 8 million links when considering only inter-domain links. This collection has 10,662 labeled hosts, 8123 normal, 2113 spam and 426 undefined. For our experiments, we considered the undefined hosts as not being spam hosts. We then considered all pages belonging to a given host as having the same label as that host. Thus all the pages of a host considered as spam are considered as spam in our experiments. Table 15 presents general statistics about this collection.

Table 15
Statistics about the WEBSPAM-UK2006 collection.

No. of pages	77,741,046	No. of hosts	11,402
No. of domains	7650	No. of links	2,951,370,103
No. of hyperarcs (host)	11,751,637	No. of hyperarcs (domain)	8,056,314

7.2. Experimental results

In the following experiments, we show the behavior of the hypergraph model regarding spam pages. We have performed experiments using the WEBSPAM-UK2006 to calculate the reputation value of each page in the database for each of the methods, computing, for each page, how many spam pages have reputation values higher than it. The list of pages was then sorted in decreasing order of reputation value and the results are presented graphically.

The use of techniques to artificially boost specific pages in search engine ranking results is nowadays commonplace on the web. This can be done in a variety of ways, including the addition of artificial linkage information in order to mislead link analysis algorithms [14]. These pages are known as web spam pages [11]. Since link analysis algorithms can be affected by such techniques, it is important to know how resistant the hypergraph model versions of link analysis algorithms are to spam techniques. If, for instance, when using HyPRDom the spam pages have a better position in the ranking than when using Pagerank alone, this could indicate that the hypergraph model makes the methods more susceptible to spam. Also, if the hypergraph model penalizes spam pages, this could indicate that part of its gain over non-hypergraph methods could be a consequence of this spam penalization.

We performed the experiments on spam with only the results obtained by the methods using the domain level (IndDom, HyIndDom, PRDom, HyPRDom), because these methods obtained the best results in the search experiments. The methods were compared pairwise (IndDom vs. HyIndDom and PRDom vs. HyPRDom). Also, for each of these pairs, two different experiments were made: one showing the results for all the pages in the database that have at least one incoming link, and one showing the results for the top 100,000 pages according to each link analysis method tested. The latter one is useful to show the impact of the hypergraph model at the top of the ranking results. Those top pages are the ones considered by the methods as having the highest reputation, which makes it worthwhile to do a more in-depth analysis of the top results.

Fig. 2 shows the results considering only the top 100,000 pages ranked by each method. It can be seen that in the top levels of the rankings, HyIndDom outperforms IndDom, giving a lower number of spam pages located in the top of the rank. This result is important because those top pages are considered as the most popular by each method.

Fig. 3 depicts two curves showing the results when considering all pages that receive at least one external

[4] http://yr-bcn.es/webspam/datasets/uk2006/

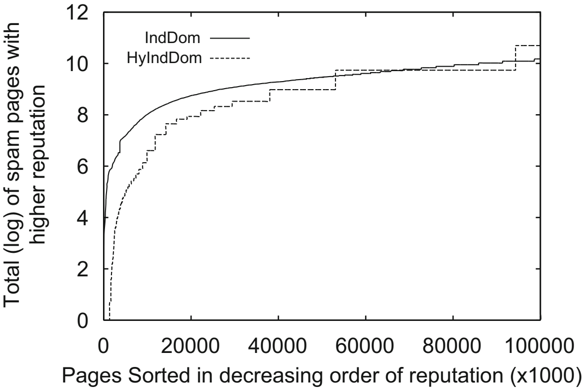

Fig. 2. Number (log) of spam pages found in a higher position than each ranking position for the first 100,000 pages.

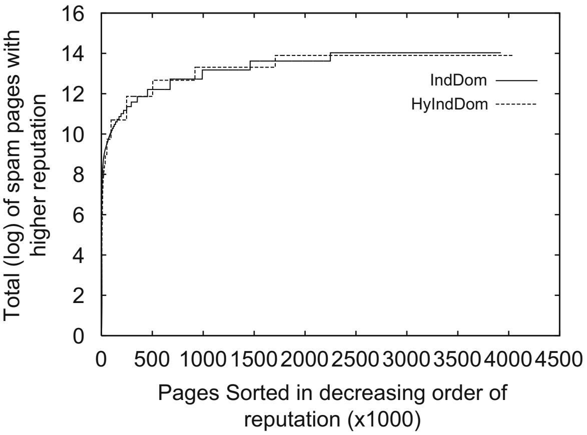

Fig. 3. Number (log) of spam pages found in a higher position than each ranking position for all pages with at least one incoming link or hyperarc.

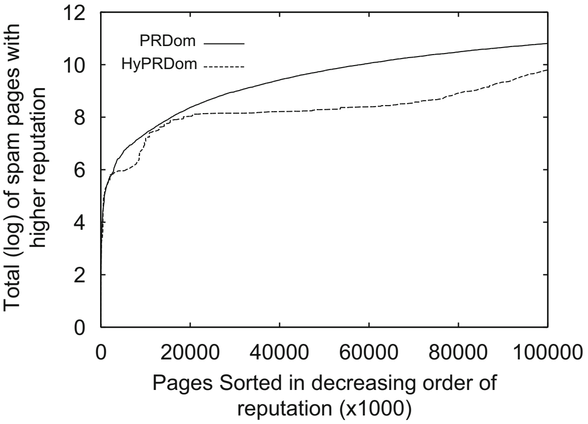

Fig. 4. Number (log) of spam pages found in a higher position than each ranking position for the first 100,000 pages.

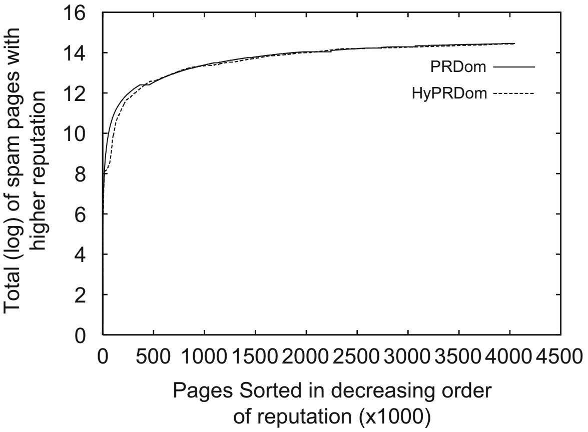

Fig. 5. Number (log) of spam pages found in a higher position than each ranking position for all pages with at least one incoming link or hyperarc.

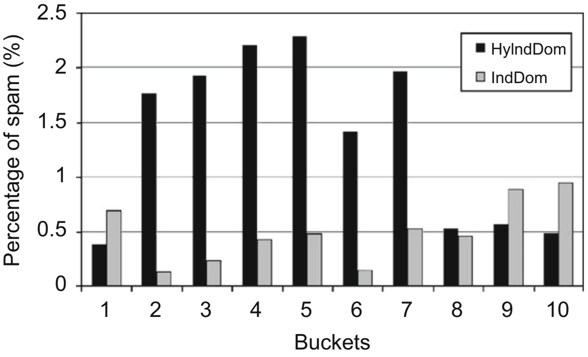

Fig. 6. Percentage of spam pages in each bucket for HyIndDom and IndDom.

link, for both IndDom and HyIndDom. These results show that the two methods have almost similar behavior towards spam pages when considering the whole set of pages. The results of the two experiments with HyIndDom show that the performance of HyIndDom was satisfactory when considering its vulnerability to spam compared to IndDom.

Figs. 4 and 5 show the results when considering only the top 100,000 pages for HyPRDom and PRDom and all pages, respectively. The results show that for practically all the ranking the HyPRDom placed the spam pages in lower positions of the ranking compared to PRDom, which implies that when used with Pagerank in the WEBSPAM-UK2006, the hypergraph model negatively affects spam pages in the database, giving them a lower position in the ranking.

We also did the same experiments performed on [21]. We sorted the pages into 10 buckets according to their reputation values, such that the sum of the reputation values of the pages in each bucket is equal to 10% of the total reputation value for all the web pages. We organized them in decreasing order, i.e., the nodes in bucket 1 have the highest reputation values, and the nodes in bucket 10 have the lowest reputation values. Figs. 6 and 7 show the distribution of the number of pages within each bucket. The dark bar represents distribution for the hypergraph methods; the light bar represents the distribution for the graph methods.

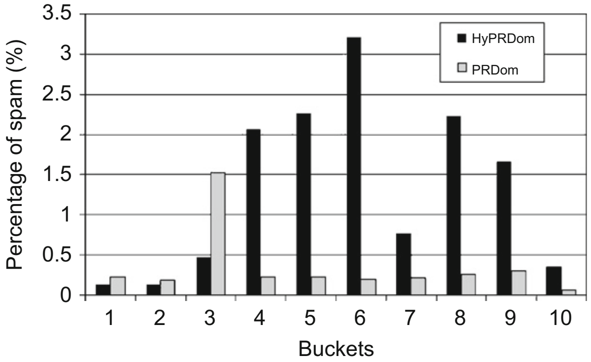

Fig. 7. Percentage of spam pages on each bucket for the HyPRDom and PRDom.

In Fig. 6, it is possible to see that the first bucket of HyIndDom contains fewer spam pages than the first bucket of IndDom. It shows that for the pages with the best reputation, the hypergraph method performs better, but only on the first 10%.

Fig. 7 shows the difference on each bucket between HyPRDom and PRDom. The performance of the hypergraph version of Pagerank is better than the graph version on the first three buckets. The set of the 30% most important pages of HyPRDom contains many fewer spam pages than the same set for PRDom.

The results shown above indicate that the use of the hypergraph instead of the original graph can lead to a quality boost regarding spam pages. However, it is important to note that this gain is not substantial, since the difference in the curves is small. It is possible to see that this gain is higher on the pages with better reputation of each method.

8. Conclusions and future work

The key advantage of using a hypergraph model of the web for computing page reputation is to allow the model to control the quality of web connections represented. This control is achieved by properly defining the partition criterion used to create hyperarcs. This flexibility opens an opportunity for further studies on determining better partition criteria, and allows search engine designers to choose the best hypergraph abstraction for their target collection.

The experiments we conducted have shown how the hypergraph model can be used to provide a better estimate of page reputation and improve the final search engine rankings. We have studied three distinct partition methods derived from the URL hierarchy: page-based, host-based and domain-based. We also have shown examples of how to adapt previously proposed link analysis methods to the hypergraph model. In the experiments performed with a real case search engine database, the WBR03 collection, the link analysis algorithms using hypergraph model provided better results for navigational queries than those obtained by using the traditional graph model, with no loss for informational queries. This is an example of the possible advantages of using the hypergraph model.

A drawback of using the hypergraph models is that the number of hyperarcs tends to be smaller than the number of arcs in the collection. Thus, in collections where there is not much connectivity information the abstraction of representing the web as a hypergraph may produce sparse hypergraphs, with few connections between the elements. In these situations the hypergraph model might cause a loss in the quality of search results. However, in experiments performed in a collection with a low degree of connectivity information, WT10g, the use of the hypergraph model in search tasks resulted in performance similar to the results obtained when representing the web as a graph.

Finally, we performed experiments regarding the effects on the reputation of spam pages caused by the adoption of the hypergraph model with domain as the partition criterion. The results indicate that the adoption of this hypergraph model caused a slight reduction in the importance given to spam pages by two page reputation algorithms, Indegree and Pagerank.

In future research, we intend to investigate further the possibility of using clustering algorithms as partition strategies. The idea is to determine the partitions based on the desired properties of the hypergraph, such as content and relationship independence between pages, instead of using only the URL hierarchy as a guide. Another direction is to study the possible correlations between the best partition criterion and the collection characteristics, using other available web collections to perform the experiments.

Acknowledgments

This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20080130192600. We thank the partial support given by, Project SIRIAA (Grant CNPq/CT-Amazônia 55.31/2005-9), Brazilian National Institute of Science and Technology for the Web (Grant MCT/CNPq 573871/2008-6), Project InfoWeb (Grant MCT/CNPq/CT-INFO 550874/2007-0), CNPq Grant 30.5237/02-0 (Nivio Ziviani), CNPq Grant 302209/2007-7 (Edleno S. de Moura), CNPq, Brazil (Klessius Berlt), CNPq, Brazil (André Carvalho), and Federal University of Goiânia, Goiânia, GO, Brazil (Thierson Couto).

Appendix

This appendix shows some samples of the queries that were used to evaluate each collection. In Tables 16 and 17, it is possible to see some popular and random navigational queries, respectively, while Tables 18 and 19 present some examples of the popular and random informational queries. Those queries are a subset of the set of queries used to evaluate the results on the WBR03 collection.

In Table 20 it is possible to see some of the navigational queries used to evaluate the results on the WT10g collection.

Table 16
A sample set of the popular navigational queries used in the WBR03 collection.

Query	Target URL
Abnt	http://www.abnt.org.br
Altavista	http://www.altavista.com.br
Assustador	http://www.assustador.com.br
Babado	http://babado.ig.com.br
Bacaninha	http://bacaninha.uol.com.br
Baixaki	http://www.baixaki.com.br
Banco real	http://www.bancoreal.com.br
Banespa	http://www.banespa.com.br
Bate papo uol	http://batepapo.uol.com.br
Bbb	http://bbb.globo.com
Blig	http://blig.ig.com.br
Bol	http://www.bol.com.br
Cade	http://www.cade.com.br
Caixa economica	http://www.caixa.gov.br
Cef	http://www.caixa.gov.br

Table 17
A sample set of the random navigational queries used in the WBR03 collection.

Query (Portuguese)	Target URL
Agencia nacional do petroleo	http://www.anp.gov.br
Allnet	http://www.allnet.com.br
Anvisa	http://www.anvisa.gov.br
Arremate	http://www.arremate.com.br
Assustador	http://www.assustador.com.br
Baixaki	http://www.baixaki.com.br
Bananagames	http://www.bananagames.com.br
Banco bradesco	http://www.bradesco.com.br
Banco central	http://www.bcb.gov.br
Banco itau	http://www.itau.com.br
Bol	http://www.bol.com.br
Buscaki	http://www.buscaki.com.br
Cade	http://www.cade.com.br
Caixa economica	http://www.caixa.gov.br
Capes	http://www.capes.gov.br

Table 18
A sample set of the popular informational queries used in the WBR03 collection.

Original query (Portuguese)	Translated query (English)
Adolf hitler biografia	Adolf hitler‘s biography
Astrologia	Astrology
Biblia	Bible
Biblioteca	Library
Biologia	Biology
Bob marley	Bob marley
Culinaria	Cooking
Curriculum vitae	Curriculum vitae
Doencas sexualmente transmissiveis	Sexually transmitted diseases
Empregos	Jobs
Fisica	Physics
Futebol	Soccer
Geografia	Geography
Humor	Humor
Indios	Indians

Table 19
A sample set of the random informational queries used in the WBR03 collection.

Original query (Portuguese)	Translated query (English)
Aborto	Abortion
Bate papo	Chat
Bee gees	Bee gees (music)
Dicionario	Dictionary
Genetica	Genetics
Hipnotismo	Hypnotism
Homem aranha	Spider man
Jogos	Games
Loterias	Lotteries
Magica	Magic
Papel de parede	Wallpaper
Piadas	Jokes
Poemas	Poems
Porto seguro	Porto seguro (city)
Shakespeare	Shakespeare

Table 20
A sample set of the navigational queries used in the WT10g collection.

Query	Target URL
HKUST Computer Science Dept	http://www.cse.ust.hk
English Server at Carnegie Mellon University	http://english-server.hss.cmu.edu
Haas Business School	http://haas.berkeley.edu
University of Wisconsin Lidar Group	http://www.lidar.ssec.wisc.edu
Donoho Design Group	http://www.ddg.com
Lockwood Memorial Library	http://ublib.buffalo.edu/lml
Brent Council	http://www.brent.gov.uk
Digital Realms	http://www.digital-realms.com
Law Society of New South Wales	http://www.lawsociety.com.au
SafeSurf	http://www.safesurf.com
Savers Investors League	http://www.savers.org
Calvert County Libraries	http://www.calvert.lib.md.us
Graduate Theology Union	http://www.gtu.edu
Boulder Community Network	http://bcn.boulder.co.us
American Chemical Society	http://www.acs.org

References

[1] B. Amento, L. Terveen, W. Hill, Does “authority” mean quality? predicting expert quality ratings of web documents, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 2000, pp. 296–303.

[2] R. Baeza-Yates, C. Castillo, Relating web characteristics with link based web page ranking, 2001, pp. 21–32.

[3] P. Bailey, N. Craswell, D. Hawking, Engineering a multi-purpose test collection for web retrieval experiments, Information Processing & Management 39 (2003) 853–871.

[4] K. Bharat, B.-W. Chang, M.R. Henzinger, M. Ruhl, Who links to whom: mining linkage between web sites, in: ICDM ’01: Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2001, pp. 51–58.

[5] K. Bharat, M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, August 1998, pp. 104–111.

[6] A. Borodin, G.O. Roberts, J.S. Rosenthal, P. Tsaparas, Link analysis ranking: algorithms, theory, and experiments, ACM Transactions on Internet Technology 5 (1) (2005) 231–297.

[7] T. Bray, Measuring the web, in: Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems, Elsevier Science Publishers, Amsterdam, The Netherlands, 1996, pp. 993–1005.

[8] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, in: Proceedings of the 7th International World Wide Web Conference, April 1998, pp. 107–117.

[9] P.P. Calado, E.S. de Moura, B. Ribeiro-Neto, I. Silva, N. Ziviani, Local versus global link information in the web, ACM Transactions on Information Systems (TOIS) 21 (1) (2003) 42–63.

[10] A. Carvalho, P.A. Chirita, E.S. de Moura, P. Calado, W. Nejdl, Site level noise removal for search engines, in: Proceedings of the 15th International Conference on World Wide Web, ACM Press, New York, NY, USA, 2006, pp. 73–82.

[11] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna, A reference collection for web spam, SIGIR Forum 40 (2) (2006) 11–24.

[12] N. Craswell, S. Robertson, H. Zaragoza, M. Taylor, Relevance weighting for query independent evidence, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 2005, pp. 416–423.

[13] C. Gurrin, A.F. Smeaton, Replicating web structure in small-scale test collections, Informational Retrieval 7 (3–4) (2004) 239–263.

[14] Z. Gyöngyi, H. Garcia-Molina, Link spam alliances, in: Proceedings of the 31st International Conference on Very Large Data Bases, 2005, pp. 517–528.

[15] D. Hawking, E. Voorhees, N. Craswell, P. Bailey, Overview of the trec8 web track, in: 8th Text REtrieval Conference, 1999.

[16] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, California, USA, January 1998, pp. 668–677.

[17] A.N. Langville, C.D. Meyer, Deeper inside pagerank, Internet Mathematics 1 (3) (2003) 335–380.

[18] R. Lempel, S. Moran, Salsa: the stochastic approach for link-structure analysis, ACM Transactions on Information Systems 19 (2) (2001) 131–160.

[19] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, first ed., McGraw-Hill, New York, 1983.

[20] M. Sanderson, J. Zobel, Information retrieval system evaluation: effort, sensitivity, and reliability, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 2005, pp. 162–169.

[21] B. Wu, B. Davison, Identifying link farm spam pages, in: Proceedings of the 14th World Wide Web Conference, 2005.

[22] G. Xue, Q. Yang, H. Zeng, Y. Yu, Z. Chen, Exploiting the hierarchical structure for link analysis, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 2005, pp. 186–193.

[23] C. Zahn, Graph theoretical methods for detecting and describing gestalt clusters, IEEE Transactions on Computers 20 (1971) 68–86.