A Hypergraph Model for Computing Page Reputation on Web Collections

Klessius Berlt[1], **Edleno Silva de Moura**[1], **André Carvalho**[1]

Marco Cristo[2], **Nivio Ziviani**[3], **Thierson Couto**[3]

[1]Departamento de Ciência da Computação – Universidade Federal do Amazonas(UFAM)
Manaus – Am – Brazil

{edleno,klessius,andre}@dcc.ufam.edu.br

[2]FUCAPI - Analysis, Research, and Tech. Inovation Center
Manaus – Am – Brazil

marco.cristo@fucapi.br

[3]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais(UFMG)
Belo Horizonte – Mg – Brazil

{nivio,thierson}@dcc.ufmg.br

Abstract. *In this work we propose a representation of the web as a directed hypergraph, instead of a graph, where links can connect not only pairs of pages, but also pairs of disjoint sets of pages. In our model, 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. Each hyperarc connects a block of pages to a single page and is created with the goal of providing more reliable information for link analysis methods. We used the hypergraph structure to compute the reputation of web pages by experimenting hypergraph versions of two previously proposed link analysis methods, Pagerank and Indegree. We present experiments which indicate the hypergraph versions of Pagerank and Indegree produce better results when compared to their original graph versions.*

1. Introduction

The reputation of web pages is one of the most successful sources of information for processing queries on web search engines. Modern search engines use algorithms that analyze the web graph for estimating the reputation of each page and then use this estimation as an additional relevance evidence when processing queries. Such strategy, known as link analysis, plays an important role in the quality of the ranking provided by search engines nowadays.

Many link analysis strategies have been proposed in literature in the last decade [3, 4, 11, 10, 13]. In all the cases, there is a common and central idea 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 such source of information was the Indegree, which uses the number of links to a page as an estimation of its reputation [3]. This first idea was then followed by sophisticated strategies, with the Pagerank [4] being one of the most known and successful of them. All these approaches adopt a graph as a model for the web, where vertices represent pages and edges represent the links between pages. Also, there is a common and central idea that a link from one page to another may represent a vote for the reputation of the destination page.

One of the main problems endured by link analysis strategies is to determine whether a link to a page can be considered as a vote for quality or not. Some links, such as navigational purpose links or spam links are noisy information in the web graph and can lead link analysis methods to wrong conclusions about the page reputation, when these links are considered as votes.

In this work we propose a representation of the web as a hypergraph, instead of a simple graph, that aims at reducing the impact of non-votes links when computing page reputation. We show how to adapt the Pagerank method [4] and the Indegree methods [3] for computing 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 $\mathcal{B}$ and a page p, there is an hyperarc from $\mathcal{B}$ to p if and only if there is one or more web links from pages of $\mathcal{B}$ to p.

The criterion adopted to group pages in a block can vary according to the final goal of the link analysis method. We here investigate the impact of different partition criteria in the task of ranking search engine answers. The main goal is to find partition criteria that allow the reduction of the negative impact of non-votes links, but keep enough information to allow links analysis algorithms to compute page reputation.

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 of our approach when compared to the traditional web graph representation is that our model aims at catching connections that are more likely to be better votes for quality than the links in the web graph. The main characteristic of the hypergraph model is that it allows controlling the influence of the individual page connections on a vote. The more fine-grained are the page blocks, the more 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 of quality of the links.

Links that do not represent votes for quality, such as Spam links and navigational links, are a problem to link analysis strategies since their importance has to be discounted when calculating page reputation. The most common approach is to identify the quality of a link by considering the importance of the page that the link comes

from. This approach is adopted by methods such as Pagerank [4] and HITS [10], where recursive processes compute the most important pages based on the importance of the pages that link to them. In spite of minimizing the problem, these strategies can still be easily affected by noise on the web [7].

The hypergraph model allows for obtaining reliable votes for quality by choosing the most appropriate partition granularity for a given collection. For instance, if links of a collection are noisy, coarse-grained blocks of pages can be used to diminish the influence of individual links on the computation of the page reputation. If the links are reliable, the simple use of these links as hyperarcs pointing to pages is also reliable.

The choice of an appropriate partition leads to high quality link analysis strategies that are useful for systems that search on the web. In experiments with a web collection the hypergraph versions of the link analysis methods produced a significant improvement on the ranking quality for navigational queries, while they maintained the ranking quality for informational queries similar to the ones generated by the graph based versions. We experimented the methods combining the page reputation with the textual content of the pages and with the anchor text information available on the collection adopted in the experiments. Two combination strategies previously proposed in literature were adopted. The experiments were performed dividing the query sets according to their types, into navigational and informational, and according to their popularity into popular and randomly selected.

Examples of the results are the comparison of the best results of Pagerank implementation using the hypergraph model and the best results of Pagerank implementation using the graph model. In this case, the hypergraph version obtained a gain of 27.8% in MRR results for popular navigational queries, 9.36% in MRR for randomly selected navigational queries and resulted in no loss when processing informational queries.

This work is organized as follows. Section 2 discusses the related work. Section 3 describes the proposed hypergraph model and new versions of Pagerank and Indegree, using the hypergraph model. Section 4 presents how to adapt Pagerank and Indegree to the hypergraph model. Section 5 presents an evaluation of our proposal, comparing the hypergraph versions of Pagerank and Indegree to their original graph versions. Finally, section 6 presents conclusions and future research directions on the topic.

2. Related work

The efforts in analyzing the link structure of the Web and using it as a source of relevance evidence in search engines started in 1996 with the Indegree method [3]. That work proposed the use of the number of incoming links of a page as a heuristic for determining the importance of each page on the Web. The intuition behind this first idea was that pages with more incoming links have more visibility and thus may have also high reputation in terms of quality.

The Pagerank method [4] is one of the most successful methods for link analysis. It computes the web page reputation scores estimating the probability of a random surfer reaching that page. An important characteristic of the Pagerank method over the Indegree method is that a web page may have high reputation without having a high number of links pointing to it, since links from pages that have high Pagerank scores have high influence in 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 in the final results than pages with low Pagerank scores. However, such Pagerank characteristic may originate a bias, since a page can receive a high Pagerank score even though it is pointed by a small number of pages [7].

The problem of having a few pages highly influencing the final scores of other pages also appear in other important and popular link analysis methods, such as the HITS method proposed by Kleinberg [10], its variant proposed by Bharat and Henzinger [2], and the SALSA method [11], which is inspired on HITS and Pagerank.

Amento et al [1] presented experiments for a site level version of Indegree, where all the links to a web site were computed as links to its root page. The idea of computing the reputation considering high level entities, such as sites or domains, is in fact explored in many previous works [3, 8]. However, as far as we know, no previous work have presented a model that allows representing both pages and high level entities together as we do here. One of the advantages of such 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 influences.

3. The hypergraph model

Our model uses a directed hypergraph to represent the web. A directed hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices $\mathcal{V}$ and a set of hyperarcs $\mathcal{E}$, where $\mathcal{E} \subseteq 2^{\mathcal{V}} \times 2^{\mathcal{V}}$. Since in our model we intend to calculate the importance of individual pages, we redefine $\mathcal{E}$ in a more restrictive way, that is, $\mathcal{E} \subseteq 2^{\mathcal{V}} \times \mathcal{V}$. Thus the hyperarcs always point to single vertices. We also require that, for each hyperarc $\epsilon = (G, v) \in \mathcal{E}$, $v \notin G$ where $G \subseteq 2^{\mathcal{V}}$. To model the web, we consider each page as a vertex of the graph and we partition the set of pages, generating non-overlapping page blocks where the pages are grouped according to an affinity criterion. We consider that a partition block $\mathcal{B}$ in the hypergraph points to a page v through a hyperarc $\epsilon = (\mathcal{B}, v)$ if, and only if, there is at least one page of $\mathcal{B}$ that has a web link to page v and $v \notin \mathcal{B}$. An important difference from this model to the traditional web graph model is that the partition criterion determines the granularity of the hyperarcs.

3.1. Partition criteria

In this work we investigate three partition criteria:

- *Page-based partition* – a block is composed by a single web page. This criterion allows for the traditional representation of the web.
- *Domain-based partition* – all the pages of a block belong to a same web domain.
- *Host-based partition* – all the pages of a block belong to a 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 studied methods.

We then adopted domain-based and host-based partitions because they group pages that are probably created by the same author or by people related to each other. This possibility was mentioned in the literature [3, 8] 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 it is for pages. As a result, we expect to obtain models where the set of partition blocks is highly reliable since the page reputation computed considering hyperarcs coming from a domain or a host is proportional to the *diversity* of partition blocks that point to the page and not to the number of links to the page, as is the case of the traditional representation of the web graph.

The 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 on 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 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 partition criteria are considered, as illustrated in Figure 1. In Figure 1(a), we show the page-based partition (simulating the traditional web graph representation). In Figure 1(b), we show the host-based partition, which groups pages with the same host name in one

block. In Figure 1(c), we consider 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 granularity creates a trade-off between the number of hyperarcs (and thus the coverage on the amount of information about each page) and the qualitative information provided by each hyperarc.

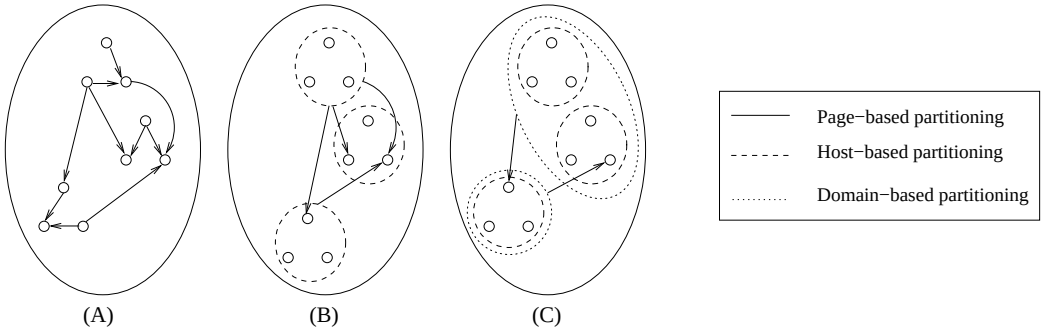

Figure 1. The hypergraph model considering (a) pages as the partition criterion, (b) hosts as the partition criterion and (c) domains as the partition criterion.

4. Computing reputation in the Hypergraph Model

To illustrate the advantages of our model, we show here examples of how two traditional link analysis methods, Pagerank and Indegree, can be adapted to our hypergraph model. Notice that other variations of these methods can be derived for the hypergraph model. These two examples are proposed to show the potential advantages of using the hypergraph model to provide a better representation of the web connections. The two adaptations will be referred to as HyperPagerank (a possible hypergraph version of Pagerank) and Hyperindegree (a possible hypergraph version of Indegree). In the next section, we present experiments to evaluate the performance of these two methods when compared to their original versions. Further studies on how to adapt these two methods and other link analysis methods to the hypergraph model will be addressed in future work.

To compute the HyperPagerank we need to propose a way to compute 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 $GR(\mathcal{B})$ is computed as the sum of the reputation of each page p that belongs to it, as

$$GR(\mathcal{B}) = \sum_{p \in \mathcal{B}} PR(p), \tag{1}$$

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

2. The pagerank value of each page depends on its representation on the hypergraph. Pages with no incoming hyperarcs have value 0, meaning that we consider that they have no reputation. For each page p with incoming hyperarcs, we give an initial value $1/\|\mathcal{V}\|$, and compute the reputation $PR(p)$ as:

$$PR(p) = (1-c) \times \sum_{\mathcal{B} \in I(p)} \frac{GR(\mathcal{B})}{\|O(\mathcal{B})\|} + \frac{c}{\|\mathcal{V}\|} \quad (2)$$

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

These two steps are repeated until the convergence of values. Notice that, as in the original Pagerank, the convergence is assured.

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:

$$HI(p) = \sum_{\mathcal{B} \in I(p)} 1, \quad (3)$$

where I(p) is defined as in Eq. (2).

5. Experiments

In this section we describe the environment setup for the experiments, discuss the studied link analysis methods and present the experimental results comparing our methods to their corresponding graph based versions.

5.1. Environment Setup

We adopted for the experiments the WBR03 collection, 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 139,402,245 links valid between its pages and the average size of plain text of each document is 5Kb. This number of valid links indicates WBR03 has a highly connected set of pages, providing reach 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 WBR03 collection.

In the experiments we have used queries extracted from a log of 3 million queries submitted to TodoBR in order to evaluate the impact of our methods within

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

Number of Pages	12,020,513	Number of Hosts	999,522
Number of Domains	141,284	Number of Links	139,402,245
Number of Hyperarcs (Host)	32,414,004	Number of Hyperarcs (Domain)	1,906,879
Average Plain Text Size	5kb		

Table 1. Statistics about WBR03 collection.

practical situations. We divided the query set into two main groups: (1) *navigational queries*, where the user is searching for a specific web site, and (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 50 most popular navigational queries found in the log. The set of randomly selected navigational queries was composed of 50 queries randomly selected from the log. For all navigational queries the results were evaluated using MRR (Mean Reciprocal Ranking), which is the metric adopted for navigational queries on the TREC Conference[2] and is the most common metric for evaluating the quality of results in navigational queries.

The set of popular informational queries contained the 50 most popular queries found in the log. The set of randomly selected informational queries was composed of 50 queries. We evaluated informational queries using the same pooling method used within the TREC web collection [9]. We thus constructed query pools containing the first 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. We then experimented with the methods in two scenarios: In the first, no other piece of evidence was applied for computing the ranking, thus avoiding interferences from evidence combination in the comparison results (these results will be referred to as *not combined*). In the second scenario, we combined the link analysis method with the result of the vector space model [12] over the textual content of the web pages and over the anchor text information, where each page is represented by the concatenation of all anchor text found in links that point to it.

Two combination approaches were experimented. One adopted a combination using a Bayesian belief network framework, as it is described in [5], and will be referred to as **BNC**. Another one adopted a brute force training-based combination

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

method described in [6], and will be referred to as **BFC**. This last scenario is useful to provide a better idea about the impact of the methods in a practical situation. For this last scenario, we also experimented with the combination without link analysis in order to assert the impact of the link analysis methods in the final ranking. In all the experiments we adopted t-test to evaluate the statistical significance of the results achieved. To provide information about the individual impact of each experimented method on the final ranking quality, we also present the results of the methods with no combination at all, which is referred to as **no combination**. However, it is important to notice that page reputation is a query independent information, and thus a ranking with no combination does not make sense. Results with **no combination** are provided just for comparison of the relative impact of each method in each query set.

5.2. Implemented Methods

To compare the results of using the hypergraph model with results found in the literature, we implemented two link analysis methods previously proposed: Pagerank and Indegree. We also implemented a "domain-based HyperPagerank" (referred to as HyPRDom), a "host-based HyperPagerank" (referred to as HyPRHost), a "domain-based HyperIndegree" (referred to as HyIndDom), and a "domain-based HyperIndegree" (referred to as HyIndHost).

The *Pagerank* was chosen for being considered a successful link analysis method. It is also usually adopted as a baseline for link analysis in literature. It computes the reputation of a page as the probability of a random surfer visiting that page. Given a page p, the Pagerank formula is:

$$PR(p) = (1 - c) \times \sum_{q \in I(p)} \frac{PR(q)}{\|O(q)\|} + \frac{c}{\|\mathcal{V}\|} \tag{4}$$

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

The Indegree consists of counting, for each page p, the number of incoming links to p. While this is a quite 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.

Since the hypergraph approaches using host and domain names naturally unconsider 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: one is the original page-based version considering all links (referred to as Pagerank and Indegree), another is a host-based version considering only links between pages in distinct hosts (referred to as PRHost and IndHost), and a third is a domain-based version considering only links between pages in distinct domains (referred to as PRDom and IndDom). Such variations are useful to check whether the improvements achieved using the hypergraph model are due to the internal links removal or not.

5.3. Experimental Results

The experimental results are separated in two distinct parts. First, we present experiments using the WBR03 collection with four distinct types of query: popular navigational queries, randomly selected navigational queries, popular informational queries, and randomly selected informational queries.

Tables 2 and 3 present the MRR results when processing the set of popular navigational queries with the Pagerank versions and the Indegree versions, respectively. Notice that in all the tables the values indicated as **no combination** are provided just to allow a comparison of the impact of the methods in the query set, since page reputation is a query independent evidence, and thus a ranking using just it does not make sense. 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 t-test and all the differences between the hypergraph versions of Pagerank and Indegree and their corresponding graph versions are significant.

Pagerank Versions (popular navigational queries)			
Method	**no combination**	BNC	BFC
Pagerank	0.2727	0.4399	0.4970
PRHost	0.2919	0.4336	0.5054
PRDom	0.3835	0.5237	0.6138
HyPRHost	0.5962	0.6032	0.6825
HyPRDom	0.6630	0.7099	0.7847

Table 2. 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)

Indegree Versions (popular navigational queries)			
Method	**no combination**	BNC	BFC
Indegree	0.5092	0.5661	0.6500
IndHost	0.5273	0.5721	0.5895
IndDom	0.6659	0.694	0.7565
HyIndHost	0.5962	0.6094	0.6885
HyIndDom	0.7494	0.7881	0.8456

Table 3. 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)

In all cases for popular navigational queries, the results were improved when considering only external links, considering as external the links between domains in domain-based methods and the links between hosts in 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 is shown in the third position of the Indegree versions in Table 3. The Hyperindegree 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 HyPRDomain and between IndDom and HyIndDom are useful to show that the improvements are not only due to removal of internal links in the hypergraph, but a consequence of the better representation of connections provided by the hypergraph model. For instance, the PRDom was the best Pagerank implementation when modeling the web as a graph for navigational queries, which indicates that the removal of internal links is positive for navigational queries in the WBR03 collection. However, for popular navigational queries, the HyPRDomain method achieved a gain of 71% when compared to PagerankDomain when using only link analysis, 49% when using BNC to combine other pieces of evidence, and 27.8% when using BFC. These results indicate its performance is not only due to the removal of internal links, but a consequence of the better representation of relationships between pages given by the hypergraph model.

Tables 2 and 3 also show that results obtained when considering domains as the partition criterion were superior when compared to the results when using hosts. This same conclusion is also obtained when experimenting 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 partition elements connected due to 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 results when using hosts as the partition criterion. When using domain, this type of noise information is reduced.

Notice that in all cases the differences between the methods are attenuated when the ranking function uses other evidences, however, the hypergraph versions of the link analysis strategies still result in improvements when compared to their original versions over the web graph. The results with the popular navigational query set indicates the proposed model is specially useful for this type of query. Thus, such information could be used to improve results in a search system that automatically classifies the queries submitted to it according to their popularity.

Tables 4 and 5 present the results obtained when experimenting the methods with the set of randomly selected navigational queries. In this case the impact of

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

varying the method for computing page reputation is smaller. Further, the advantages of removing external links are attenuated, which is explained by the smaller number of external links pointing to the searched pages in this query set.

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 it can be seem in Table 4, 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 5, the results are quite close, with a slight advantage for the hypergraph when using BNC combination and a slight advantage for the graph versions when using BFC. Considering no combination, the best results were obtained by the hypergraph versions of Indegree.

Pagerank Versions (random navigational queries)			
Method	**no combination**	BNC	BFC
Pagerank	0.2911	0.5939	0.5878
PRHost	0.3428	0.5122	0.586
PRDom	0.4564	0.6145	0.7041
HyPRHost	0.5538	0.6445	0.7221
HyPRDom	0.5849	0.6982	0.7700

Table 4. 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)

Indegree Versions (random navigational queries))			
Method	**no combination**	BNC	BFC
Indegree	0.4383	0.6301	0.6924
IndHost	0.4524	0.5596	0.7604
IndDom	0.5353	0.6495	0.8219
HyIndHost	0.5286	0.5838	0.7484
HyIndDom	0.6576	0.6972	0.8152

Table 5. 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)

Tables 7 and 6 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 navigational queries because in this case users are more interested in the content of the answer pages, which makes the page content more important than in navigational queries. Thus, a lower variation in the

Pagerank Versions (popular informational queries)						
Method	**no comb.**		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

Table 6. 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)

Indegree Versions (popular informational queries)						
Method	**no comb.**		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 7. 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)

quality of results is expected when changing only link analysis strategies for this query type.

Another important detail is that the training performed by BFC resulted in a function that gives 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 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 to not show the tables for non-popular informational queries to avoid too much repetition.

In summary, the experiments presented indicate that HyPRdom can be considered as a very good alternative for link analysis, being the best option for navigational queries and giving performance equal to the remaining methods in informational queries.

6. Conclusions and Future Work

The key advantage of using a hypergraph model of the web for computing page reputation is to allow a modeling that controls the quality of web connections represented.

This control is achieved by properly defining the partition criterion adopted 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 the target collection.

The experiments we conducted have shown how the hypergraph model can be used to provide a better estimation of page reputation and improve the final search engine ranking. As examples of partition criteria, we have presented a study with three distinct kinds of partitions: 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, WBR03, the link analysis algorithms using hypergraph model presented better results for navigational than those obtained by their using traditional graph model, with no loss for informational queries. This is an example of the possible advantages of using the hypergraph model.

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

References

[1] B. Amento, L. Terveen, and 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*, pages 296–303, New York, NY, USA, 2000. ACM Press.

[2] K. Bharat and 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*, pages 104–111, Melbourne, Australia, August 1998.

[3] T. Bray. Measuring the web. In *Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems*, pages 993–1005, Amsterdam, The Netherlands, The Netherlands, 1996. Elsevier Science Publishers B. V.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, April 1998.

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

[6] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM Press.

[7] A. L. da Costa Carvalho, P. A. Chirita, E. S. de Moura, P. Calado, and W. Nejdl. Site level noise removal for search engines. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 73–82, New York, NY, USA, 2006. ACM Press.

[8] C. Gurrin and A. F. Smeaton. Replicating web structure in small-scale test collections. *Inf. Retr.*, 7(3-4):239–263, 2004.

[9] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the trec8 web track. In *8th Text REtrieval Conference*, 1999.

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

[11] R. Lempel and S. Moran. Salsa: The stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001.

[12] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1st edition, 1983.

[13] G. Xue, Q. Yang, H. Zeng, Y. Yu, and 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*, pages 186–193, New York, NY, USA, 2005. ACM Press.