Webology, Volume 4, Number 3, September, 2007

Increase of Precision on the Top of the List of Retrieved Web Documents Using Global and Local Link Analysis

Luiz Fernando de Barros Campos
School of Information Science, Federal University of Minas Gerais, Belo Horizonte, Brazil. E-mail: lfbcampos (at) gmail.com

Received January 17, 2007; Accepted February 6, 2007

Abstract

At present, information derived from the cross-references among pages is used to improve the results of Web-based information retrieval systems, as constantly occur in bibliometric techniques. The references are local when only the links related to the set of documents returned as answers to a user query are treated, as done by the HITS algorithm. If all the links of the documents in the collection are taken into account, we speak of global references. This is the case with the PageRank algorithm, which takes advantage of the whole Web structure. Using the WBR99 reference collection, the article shows the results of the implementation of the HITS and PageRank algorithms and emphasizes the gains in precision on the top of the list compared with the results of the space vector model algorithm (SVM), which is grounded only on the textual analysis of the pages. It was noticed that the use of local links produces higher average precision. However, the use of global links is justified whenever high precision at low recall is important and query processing efficiency is essential, such as in Web search engines.

Keywords

Link analysis; HITS; PageRank; Space Vector Model; Search engines

Introduction

Information derived from cross-references has been constantly utilized in bibliometrics for determining the impact factor of scientific journals, accomplishing bibliographic coupling or co-citation analysis (Small, 1973), measuring document importance (Garfield, 1972), and mixing citations with keywords aiming at improving document retrieval (Salton, 1971), among other uses.

Recently, information retrieval systems in the Web have been employing cross-references information. Usually, these systems show better results when data about the hyperlink structure of Web pages is taken into account. Classical ordering algorithms, like the *Spatial Vector Model* (SVM), were improved by considering the Web pages that point to the pages contained in the answer set, as well as the Web pages that the pages in the answer set point to. Many implementations use the number of *inlinks* and *outlinks* for better ordering and, frequently, consider a measure of importance of the referring and referred Web pages. The

basic principle is that the existence of a hyperlink to a Web page constitutes a measure of its popularity, importance or quality (i.e., a pretension of the page relevance).

When they were in Stanford, creating the bases of the Google enterprise, the pioneers Sergey and Brin created a ranking algorithm based on the graph of the links formed by the Web pages (Brin & Page, 1998; Page et al., 1998). It was called PageRank and used a popularity measure to rank Web pages. Kleinberg (1998) has specified HITS, an algorithm that classifies the pages as authorities or hubs. A good hub Web page points to many authority Web pages. Conversely, a good authority Web page is pointed to by many hub Web pages.

Kleinberg's (1998) proposal is restricted to the Web pages contained in the answer set returned to a user query, together with other Web pages that refer to them or are referred to by them. On the contrary, the PageRank algorithm proposed by Brin and Page (1998) includes all the Web pages (ideally). Based on this opposition, it is professed in this work that local link information refers to information provided by the documents contained in the answer set returned to the user query, and global link information is extracted from all Web pages, following Calado et al. (2003) and Xu and Croft (1996).

Using the *WBR99* collection, this article intends to show, detail and compare the results of the implementation of the SVM, PageRank and HITS algorithms. The SVM was adopted as the evaluation basis, since it is grounded only on textual analysis of Web pages. The most relevant documents were obtained for each one of the 50 queries contained in the reference collection and the results were evaluated by means of the precision graphics on the usual eleven recall levels and the *average precision at seen relevant documents*. Considering that the average user rarely examines documents beyond the ten first ones retrieved, improvement on the top of the list of retrieved pages is desirable. So, this article aims to obtain clues to answer the two following questions related to the ranking of documents on the top of the list retrieved as an answer to a user's query in the Web: (a) Does the link analysis effectively improve the answer compared to the textual analysis of the pages? (This is answered by comparing the results of the link analysis algorithms - PageRank and HITS - and the SVM); and (b) How does the use of global and local links information influence the results of the link analysis? (This is answered by comparing the results of the PageRank and HITS algorithms).

Description of the algorithms

The results of the SVM algorithm are taken as the reference basis, since its efficiency is mostly acknowledged and its application is generalized in information retrieval systems. For this implementation, the description of the algorithm is obtained from Salton and McGill (1983) and Baeza-Yates and Ribeiro-Neto (1999). The similarity level between each document and the user query (value used for ranking the retrieved documents) is given by:

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

The documents *j* and the user query are represented by vectors (respectively, $\vec{d}j$ and $\vec{q}$). The weights are the commonly used term frequency (*tf*) and inverse document frequency (*idf*). The term frequency is how many times the term appears in the document (sometimes

it is normalized, dividing by the maximum term frequency for the document). The inverse document frequency is calculated as follows:

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

where N is the total number of documents in the collection and n_i is the number of documents in which the term i appears. The most usual formula employed to calculate the weight of the terms in the documents is:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{n_i}\right)$$

In the case of the user queries, the weight of each term takes into account only its inverse document frequency.

The PageRank algorithm is described by Page et al. (1998). Haveliwala (1999) presents and compares many implementations of PageRank. However, the version of the algorithm described in Kamvar et al. (2003a) will be used, detailed as follows. ***P*** is the transitional stochastic matrix obtained from the Web structure and built starting from the graph G which represents the connections between all the n Web pages. Each of its entries is zero if the page i (row) does not point to the page j (column). Otherwise, it is $P_{ij} = 1/\deg(i)$, the inverse of out degree deg(i) - the number of documents to which the page i points. The matrix ***P*** undergoes two transformations. The first one is:

$$\vec{v} = \left[\frac{1}{n}\right]_{n\times 1} \qquad D = \vec{d} \cdot \vec{v}^{T} \qquad P' = P + D$$

The total number of documents in the collection is n. The vector $\vec{v}$ corresponds to an equiprobable distribution of the documents in the collection. The vector $\vec{d}$ contains entries 1 for documents i whose out degree deg(i) equals null, and entries 0 otherwise. Thus, the resulting matrix P' contains rows whose entries are $1/n$ for documents i that satisfy deg(i) =0. The second transformation is:

$$E = [1]_{n\times 1} \times \vec{v}^{T} \qquad P'' = cP' + (1-c)E$$

The final result is that the rows corresponding to documents whose out degree is null (generated by the last transformation) remain unaltered. The other rows will have entries equal to $(1-c)/n$ in columns corresponding to documents not linked by document i and entries equal to $c/\deg(i)+(1-c)/n$ in columns corresponding to the documents linked. It is noted that the term c 'distributes' probabilities between pages linked (for which the transition is random with probability $(1-c)/n$) and pages not linked (for which the transition has probability $c/\deg(i)$). The constant c employed is 0.85, value used by many other works (for instance, Kamvar et al., 2003a, or Page et al., 1998). An intuitive interpretation for this procedure is found in Page et al. (1998) - the 'random surfer' that occasionally jumps randomly to another page not linked.

The PageRank algorithm is condensed by Kamvar et al. (2003a) in this way:

$$
\begin{aligned}
&function \quad pageRank \quad (G, \vec{x}^{(0)}, \vec{v})\{ \\
&build\ P\ de\ G: P_{ji} = 1/\deg(j); \\
&repeat \\
&\qquad \vec{x}^{(k+1)} = cP^T \vec{x}^{(k)}; \\
&\qquad w = \left\| \vec{x}^{(k)} \right\|_1 - \left\| \vec{x}^{(k+1)} \right\|_1; \\
&\qquad \vec{x}^{(k+1)} = \vec{x}^{(k+1)} + w\vec{v}; \\
&\qquad \delta = \left\| \vec{x}^{(k+1)} - \vec{x}^{(k)} \right\|_1; \\
&until\ \delta < \varepsilon; \\
&return\ \vec{x}^{(k+1)}; \\
&\}
\end{aligned}
$$

PageRank is an application of the power method. The transposed matrix P^T is a stochastic matrix, as each column represents the probability of a surfer reaching one of the *i* pages pointed to by page *j*. A condition of the convergence of the Markov chain is that the stochastic matrix be regular, that is, constituted of positive entries (not null). The essential advantage of the algorithm is that it manipulates the sparse original matrix ***P***, generating the effect of manipulating the matrix transformed by the two procedures described. The convergence conditions and the mathematical details are obtained from Anton and Busby (2006). The stopping criterion is determined comparing the norms $L1$ of the two vectors obtained in consecutive iterations. Weights $1/n$ are attributed to each page (entry) in the initial vector (this vector can be imbalanced if different initial probabilities are attributed to the pages). The final vector, which converges from any initial vector if the described conditions are fulfilled, ranks the *n* pages of the collection.

Kleinberg (1998) shows and discusses the bases of the HITS algorithm, and its implementation and applications like social nets and information retrieval in the Web. Besides, analogies to bibliometric methods are made. Grounded on this article and Anton and Busby (2006), the HITS algorithm, also an application of the power method, is generically described by the following steps:

- The adjacency matrix $P = |a_{ij}|$, where $a_{ij} = 1$ if the *i*-th page links to the *j*-th page, and $a_{ij} = 0$ otherwise, is generated from the initial set of retrieved pages.
- The matrix is expanded to include pages which link to pages in the initial set or are linked by them.
- The initial hub vector (h_0) is produced by adding the entries in the rows of *P* and the initial authority vector (a_0) is produced by adding the entries in the columns of *P*.
- The hub and authority vectors are calculated by using the formulas $h_n = P \times a_{n-1} / |P \times a_{n-1}|$ and $a_{n+1} = P^T \times h_n / |P^T \times h_n|$ till the stopping criterion is reached.

- The final arrangement of the retrieved pages is obtained from the combination of the three vectors calculated by the SVM and the HITS.

It is easy to show that the hub and authority vectors can be expressed as $h_k = (P \times P^T) \times h_{k-1} / |(P \times P^T) \times h_{k-1}|$ and $a_k = (P^T \times P) \times a_{k-1} / |(P^T \times P) \times a_{k-1}|$. Since the matrixes $P \times P^T$ and $P^T \times P$ are symmetrical, they have an eigenvalue that is strictly greater in magnitude than its other eigenvalues (i.e., a dominant eigenvalue) which guarantees the convergence of the method. The factors are normalized in order to avoid the explosion of the values. It is important to note that the calculus of the HITS depends on the initial set of retrieved pages. Therefore, it occurs in real-time, not like PageRank.

Methodology

1 The WBR99 collection

The *WBR-99* collection, described by Calado (1999), was built from a set of documents collected from the Brazilian Web in November 1999. It was taken from the database of *TodoBr*, a search engine for the Brazilian Web. It comprises 5,939,061 documents, 2,669,965 terms and 40,871,504 links. The *idf* values for each term and the *tf* values and the norms of the vectors for each document are provided. The inverted lists files are also provided.

For each HTML document, its *inlinks* (links that point to the document) and *outlinks* (outgoing links from the document) are listed. The links are internal or external. An external link is an outlink to a page on a different site or an inlink from this page. An internal link is an outlink to a page on the same site or an inlink from this page. The sites are considered to be different if they are under a different domain, where by domain is meant the first part of the URL (from 'http://' through to the first '/' symbol).

In the collection, 33,154 queries are available, which were extracted from the query log of the search engine *TodoBr*. Fifty of the queries were evaluated and, therefore, utilized to compare the algorithms in this work. For these reference queries their terms and documents in their answer set were specified. The average number of terms per query is 1.78. Among these queries, 28 are generic, like 'movies' or 'mp3'. Other 14 queries are more specific, but still focusing on a general topic, like 'electronic commerce'. Finally, the last eight queries are rather specific, consisting basically of names of rock bands. The documents in the query set were evaluated by forming pools by the top 20 documents generated from the execution of each of the 50 queries, utilizing a set of six algorithms described in Calado et al. (2003). All documents in each query pool were submitted to a manual evaluation by a team of 29 users, all of them familiar with Web searching. Users were allowed to follow links and judge the pages not only to the textual content, but also to their linked pages and graphical content. The pooling method adopted was similar to the method used for the web-based collection of TREC (Text REtrieval Conference), as described by Baeza-Yates and Ribeiro-Neto (1999).

The documents judged relevant were given one of three different values: (1) meaning that the document, although not completely relevant, contained related information or links; (2) meaning that the document contained related information or links to relevant documents; and (3) meaning that the document contained relevant information.

2 Practical aspects and criteria of the implementation

The same data structures and auxiliary routines were employed for the implementation of the three algorithms. Aiming at satisfactory performance, particularly in the case of HITS,

whose vectors are calculated in real-time, a hash table was constructed containing pointers to binary trees, thus diminishing the number of collisions. For the implementation of PageRank, an algorithm was projected to balance the initial load of the documents. The lists of the documents were maintained in internal memory during the processing.

In the case of SVM, the documents retrieved contained all the terms of the query. Retrieving documents that contained any of the terms of the query did not significantly increase the precision, but considerably increased the processing time. Besides that, tests were conducted using the internal and external links, which also increased the processing time without conspicuously improving the precision on the top of the list of retrieved documents. A possible partial explanation for this is that the main pages of the sites are usually the most linked by external links as well by internal links, which follows from the Web structure. Therefore, only the external links were considered in the implementation of PageRank and HITS. It should be noted that loss of relevant information for the ranking of the retrieved documents could occur in this procedure, especially if other variables or ways of combining the results of the link analysis algorithms are taken into account, as discussed below in *Research Approaches*.

For PageRank, the matrix $\boldsymbol{P}^T$ was built from the graph of the pages in the collection. For HITS, the adjacent matrix was built from the top 100 pages returned by the SVM to each query. The stopping criterion was reached when all the entries of the current vector did not differ more than 1% from the entries of the vector previously calculated. The final ranking was obtained combining the vectors calculated by the SVM and PageRank (or HITS).

Results

In congruence with the goals outlined, the results for the 100 first retrieved documents are shown. Precision is the fraction of the retrieved documents which are judged relevant. Recall is the fraction of the documents judged relevant which have been retrieved. To evaluate the performance of each algorithm, a precision *versus* recall curve was generated by averaging the precision figures of the 50 queries at each recall level (from 0% to 100%) and interpolating when necessary.

The *average precision at seen relevant documents* was also taken into account for the evaluation, since this measure favors systems that retrieve relevant documents early in the ranking (Baeza-Yates & Ribeiro-Neto, 1999) and fits the aim of improving precision on the top of the list. The average precision at seen relevant documents is a single value summary of the index generated by averaging the precision figures obtained after each new relevant document is retrieved.

Figures 1 and 2 show the results of the implementations of the PageRank and HITS algorithms compared to the SVM for the 50 queries. In these figures, the documents in the collection whose evaluated relevance were given values 1, 2 or 3 were retrieved. There were gains of about 10% in the 0% recall level for both the PageRank and HITS algorithms.

Figure 1. Precision *versus* Recall Curve - Comparison between the implementation of the PageRank and SVM algorithms for the 100 first retrieved documents (relevance 1, 2 and 3).

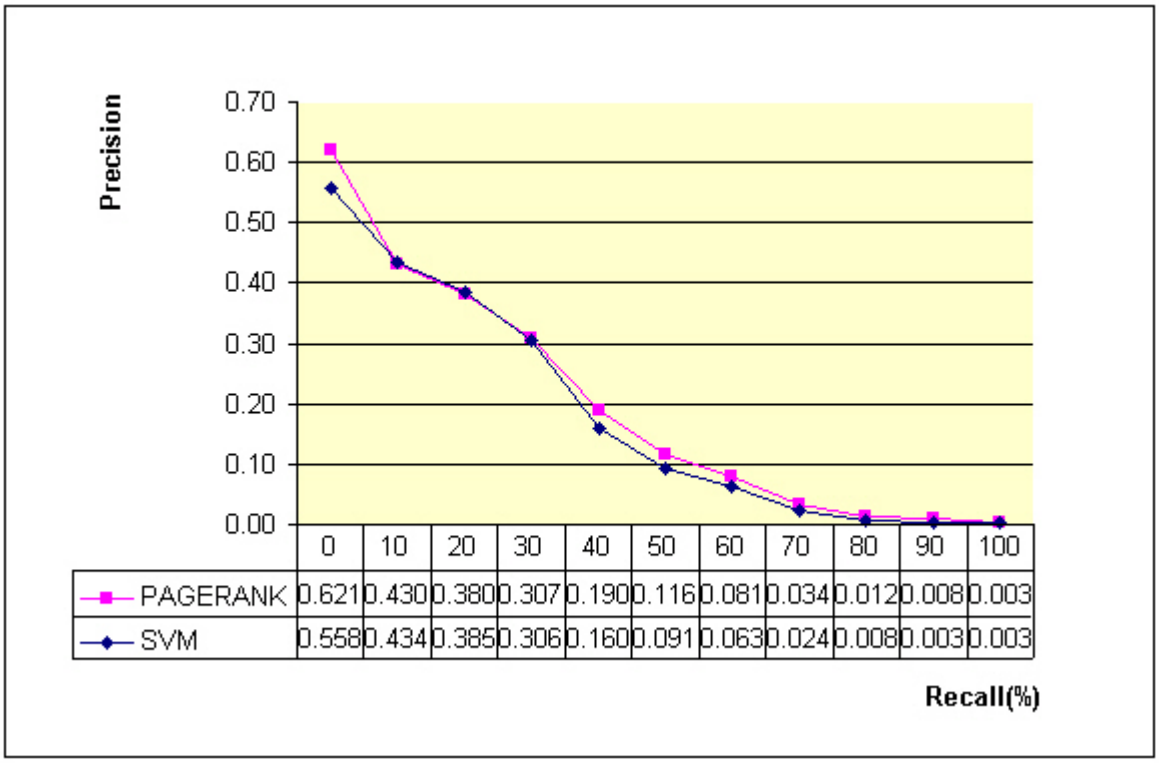


Figure 2. Precision *versus* Recall Curve - Comparison between the implementation of the HITS and SVM algorithms for the 100 first retrieved documents (relevance 1, 2 and 3).

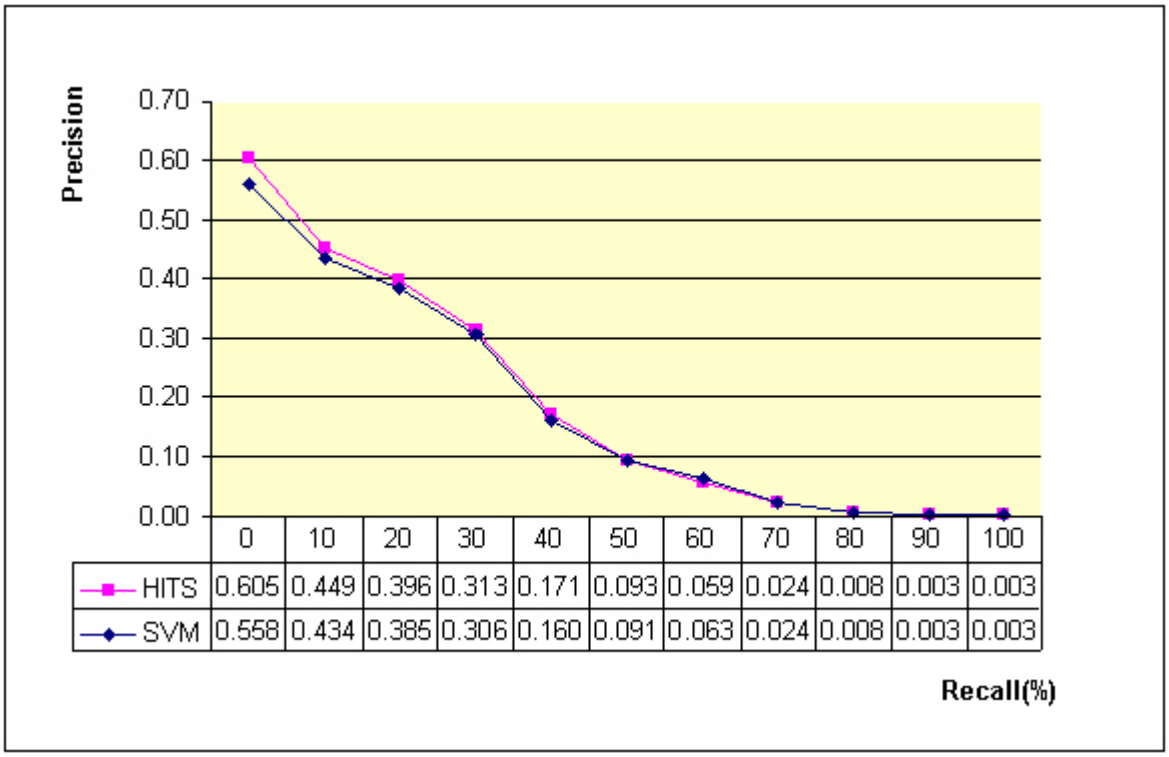
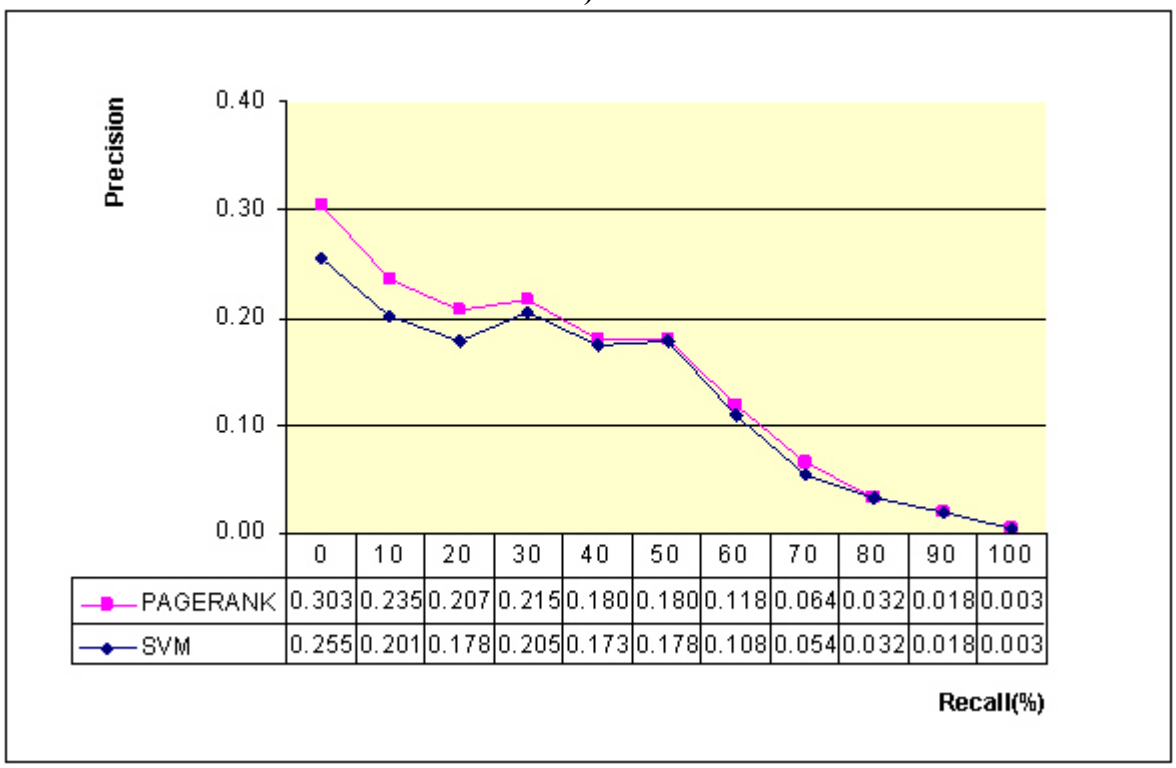


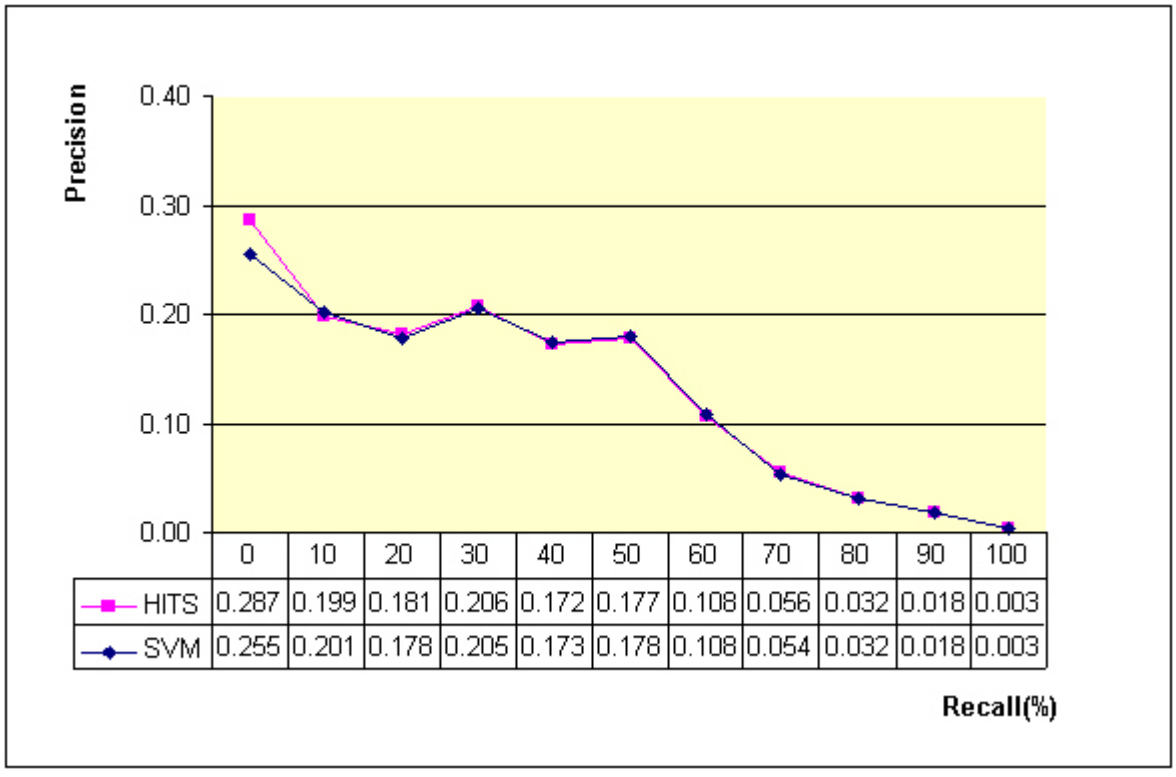
Figure 3 and 4 show the results of the implementation of the PageRank and HITS algorithms compared to the SVM for the 50 queries. In this case, only the documents in the collection whose evaluated relevance was given value 3 were retrieved. There were gains in the 0% recall level of about 19% for the PageRank algorithm and about 12% for HITS.

Figure 3. Precision *versus* Recall Curve - Comparison between the implementation of the PageRank and SVM algorithms for the 100 first retrieved documents (relevance 3).



	0	10	20	30	40	50	60	70	80	90	100
PAGERANK	0.303	0.235	0.207	0.215	0.180	0.180	0.118	0.064	0.032	0.018	0.003
SVM	0.255	0.201	0.178	0.205	0.173	0.178	0.108	0.054	0.032	0.018	0.003

Figure 4. Precision *versus* Recall Curve - Comparison between the implementation of the HITS and SVM algorithms for the 100 first retrieved documents (relevance 3).



	0	10	20	30	40	50	60	70	80	90	100
HITS	0.287	0.199	0.181	0.206	0.172	0.177	0.108	0.056	0.032	0.018	0.003
SVM	0.255	0.201	0.178	0.205	0.173	0.178	0.108	0.054	0.032	0.018	0.003

Table 1 shows the average precision at seen relevant documents (relevance 1, 2 or 3) for the three, five and 100 first retrieved documents. For the three first documents, the gains on average precision of the PageRank and HITS algorithms were in the order of 10%. For

the five first documents, the gains reached 8% for HITS and 5% for PageRank (see Table 1).

Table 1. Average precision generated for the *n* first seen relevant documents (relevance 1, 2 or 3) for the implementations of the algorithms SVM, PageRank and HITS (n=3, 5 and 100).

	Average precision at the n first seen relevant documents		
	n=3	n=5	n=100
SVM	0.432	0.403	0.360
PageRank	0.473	0.424	0.367
HITS	0.474	0.436	0.369

Conclusions

Considering the goal of improving the precision on the top of the list, the results were clearly positive and three aspects were especially congruous with this objective and deserve mention.

The first point to emphasize is that the gains in precision were higher in the first recall levels, especially the 0% level, which can be verified by simple visual inspection of the four curves illustrated. The second point refers to the higher precision gains in the 0% recall level in figures 3 and 4 compared with figures 1 and 2. This is particularly encouraging, since figures 3 and 4 refer to the documents judged strictly relevant (category 3), which were ranked considerably better. PageRank was especially efficient in ordering these documents. The third point relates to the average precision generated at the n first seen relevant documents (see Table 1), which was higher for a lower n, that is, it was higher on the top of the list.

There was evidence to state that using algorithms which compute the links led to gains in precision in relation to the space vector model (SVM) algorithm, which is focused on the documents' content, especially on the top of the list. There are works that have reached opposite conclusions and reported no gains in the ranking of the retrieved documents when analyzing the links, like Hawking and Craswell (2001). Nevertheless, the TREC collection was used in this last work. Calado et al. (2003) noted three differences between the TREC and WBR99 collections that could justify the divergence. First, the documents of the TREC collection were judged based only on their text, since following the links of the documents was not allowed to the evaluators. On the contrary, the judges had access to the real site and could follow links in the case of the WBR99, as previously commented. Second, link analysis algorithms work better with generic queries, as is the case in the Web scenario. This justifies partially the better results, since the TREC's queries are more specific. Third, the documents in the WBR99 were collected by a Web crawler, whereas TREC was constructed as a subset of a larger collection. Therefore, its link distribution is very likely quite different from the Web, which could cause disparities in the results.

The second part of the research question stated the comparison between the results of the PageRank and HITS algorithms. Interestingly, PageRank always presented better results than HITS at the 0% recall level. Nevertheless, the average precision at seen relevant documents for HITS was higher at the three measure points illustrated in Table 1. This suggests that local link analysis algorithms effectively bring about some gain in the ranking of retrieved documents, but the use of global link analysis algorithms is justified whenever high precision at low recall is required and query processing efficiency is essential, since there is no processing in real-time in the case of PageRank. These are

exactly two of the conditions which must be accomplished by Web search engines. Similar conclusions were reached by Calado et al. (2003).

Research approaches

The implementation of the algorithms and the analysis of the results have made it possible to outline related research which employs diverse approaches.

One of these approaches is the improvement of the computation of the algorithms. As the Web expands exponentially, the operations with the matrix of its page graph turned out to be critical. In the last few years, various studies have been undertaken aimed at improving or adapting the processing of the matrix, speeding it up or carrying it out under restraining conditions, as shown in the following examples. The Web structure, keeping in mind that the majority of the links point to the initial (main) pages of the sites, is exploited for processing PageRank in Kamvar et al. (2003a). Some extrapolation methods for accelerating the algorithm's convergence are discussed in Kamvar et al. (2003b). In Chen et al. (2002) some techniques for calculating PageRank in computers with limited main memory are analyzed.

Another possible research approach is the enhancement of the ranking algorithms by means of combining the document textual analysis and the link analysis. It is noteworthy that the commercial search engines take into account innumerable variables. The possibilities are countless. Weighting internal link information can be useful; mainly text information in the documents, like proximity of the query terms to the links, existence of the query terms in the hyperlinks, and many others have proved to be beneficial. Some examples follow. Chakrabarti et al. (1998) propose the mixing of the link analysis algorithm described in Kleinberg (1998) and the utilization of words located around the links to calculate their weights. Bharat and Henzinger (1998) expand the original queries using keywords obtained from the documents in the local answer set and attribute weights to the links grounded on the expanded query.

Looking forward to enhancing the processing of the algorithm, usually we run into information or situations which can power the combination of textual and link analysis. This implies applying both of the approaches explained, as in Wang and DeWitt (2004), who exploited an interesting technique, the processing of the adjacency matrix in many parts by servers in a distributed Internet search system. Their strategy can refrain from processing some parts of the matrix (considering the already computed results) or enhance the algorithm (considering that the division of the matrix implies the use of information which can be employed for the vector calculus and improvement of the algorithm). Under this view, Haveliwala (2002) builds not only one PageRank vector for the Web pages, but many vectors referring to diverse topics. Most likely, contextual information will be gradually employed to refine the ranking vector. Again, Haveliwala (2002) chose topics for determining the PageRank vectors used grounded on contextual information (words highlighted in a Web page). Besides, for a user, queries can be optimized based on patterns obtained from the analysis of previous queries. This could help prevent difficulties like synonymy or polysemy. It looks as though modern commercial search engines have already started applying this procedure.

Acknowledgments

I am grateful for the availability of the WBR-99 for academic research.

References

- Anton, H., & Busby, R. C. (2006). *Álgebra linear contemporânea*. Editora Bookman, São Paulo.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley.
- Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia), p. 104-111.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference* (WWW7).
- Calado, P. P. (1999). *The WBR-99 collection. description of the WBR-99 collection data-structures and file formats*. LATIN - LaboratÃ³rio para o Tratamento de Informação, Departamento de Computação, Universidade Federal de Minas Gerais.
- Calado, P.P., Ribeiro-Neto, B., Ziviani, N., Moura, E., & Silva, I. (2003). Local versus global link information in the Web. *ACM Transactions on Information Systems*, 1(21), 42-63.
- Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane), p. 65-74.
- Chen, Y., Gan, Q., & Suel, T. (2002). I/O efficient techniques for computing PageRank. *CIKM* 2002, p. 549-557, McLean, Virginia.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471-479.
- Haveliwala, T. H. (1999). Efficient computation of PageRank. *Stanford University Technical Report*.
- Haveliwala, T. H. (2002). Topic-sensitive PageRank. *Proceedings of the 11th International World Wide Web Conference* (WWW11).
- Hawking, D. and Craswell, N. (2001). Overview of TREC-2001 Web Track. *Proceedings of the Tenth Text REtrieval Conference* (TREC-2001), Gaithersburg, MD, p. 61-67.
- Kamvar, S.D., Haveliwala, T. H., Manning, C.D., & Golub, G.H. (2003a). Exploiting the block structure of the Web for computing PageRank. *Stanford University Technical Report*.
- Kamvar, S.D., Haveliwala, T.H., Manning, C.D., & Golub, G.H. (2003b). Extrapolation methods for accelerating PageRank computations. *Proceedings of the 12th International World Wide Web Conference* (WWW12).
- Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. *Stanford Digital Libraries Working Paper*.
- Salton, G. (1971). Automatic indexing using bibliographic citations. *Journal of Documentation*, 27(2), 98-110.
- Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. 1st ed., McGraw-Hil, New York.
- Small, H. G. (1973). Co-citation in the scientific literature: A new measure of relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.
- Wang, Y., & DeWitt D.J. (2004). Computing PageRank in a distributed Internet search system. *Proceedings of the 30th VLDB Conference, Toronto, Canada*.
- Xu, J., & Croft, W.B. (1996). Query expansion using local and global document analysis. *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich), p. 4-11.

Bibliographic information of this paper for citing:

Campos, Luiz Fernando de Barros (2007). "Increase of precision on the top of the list of retrieved web documents using global and local link analysis." *Webology*, **4**(3), Article 44. Available at: http://www.webology.org/2007/v4n3/a44.html