Impact of Contextual Information for Hypertext Documents Retrieval

Idir Chibane and Bich-Liên Doan

SUPELEC, Computer Science dpt.
Plateau de Moulon, 3 rue Joliot Curie, 91 192 Gif/Yvette, France
{Idir.Chibane, Bich-Lien.Doan}@supelec.fr

Abstract. Because the notion of context is multi-disciplinary [17], it encompasses lots of issues in Information Retrieval. In this paper, we define the context as the information surrounding one document that is conveyed via the hypertext links. We propose different measures depending on the information chosen to enrich a current document, in order to assess the impact of the contextual information on hypertext documents. Experiments were made over the TREC-9 collections and significant improvement of the precision shows the importance of taking account of the contextual information.

1 Introduction

Since the beginning of the Web, information has become widely-accessed and widely-published. The volume of heterogeneous and distributed information available on the Web has been exponentially and continuously growing. That's why the seeking and selection of relevant information is a very complex and difficult task. Search engines help the final-user in this retrieval task by indexing a part of the Web, but they have very few information concerning the information need of the user. Experiments show that most of user's requests contain 2 or 3 terms. So few numbers of terms often leads to noise and silence in the responses given by search tools. This is a consequence of several reasons that include, among others, the implicit user's information need (for example her intention, the context of the query) and the non use of contextual information of the documents in the indexing phase. Several works on survey attempted to classify different contexts alongside with functional or opposite criteria. For [14], [15] and [16], the context of a document is the information related to the current document that is conveyed through hypertext links, semantic network, or surrounding text. The context is used to enrich the local index of a document with information extracted from its neighbours. Experiments showed that taking account this context provide better precision for certain types of queries.
In this paper, we are particularly interested in the local context of Web resources and we define the **context of Web pages** as the neighbourhood information of pages

which is brought from the hypertext links to all resources directly related to these current pages. In recent years, several information retrieval methods using the information about the link structure have been developed and proved to provide significant enhancement to the performance of Web search in practice. Actually, most of systems based on link structure information combine the content with the popularity measure of the page to rank a query result. Google's PageRank[1] and Keinberg's HITS[2] are two fundamental algorithms employing the hyperlink structure among the Web page. A number of extensions to these two algorithms are also proposed, such as [3][7][8][9][10][11]. All these link analysis algorithms are based on two assumptions: (1) the links convey human endorsement. If there is a link from page A to page B, then we may assume that page A endorses and recommends the content of page B. Thus, the importance of page *A* can, in part, spread to the pages besides B it links to. (2) Pages that are co-cited by a certain page are likely to share the same topic as well as to help retrieval.
The study of the existing systems enabled us to conclude that all ranking functions based on link structure information do not depend on query terms. It decreased significantly the found results precision. Indeed, analysis of the user's behaviours in their research shows that they are not interested in the popular pages, if it does not contain the query terms. In this paper, we first review the related literature on link analysis ranking algorithms. We also present some extension of these algorithms, by defining the context of Web pages as enriched neighbourhood information conveyed through hypertext links and whose importance is computed according to the query terms. Then, we introduce our new link analysis ranking algorithm with the new ranking function and we present experiments on multiple queries, using the proposed algorithm. We also present a comparative of different link analysis ranking algorithms. Last, we discuss results' analysis.

2 Related Work

Various studies suggested that taking account of links between documents increases the quality of information retrieval. PageRank[1] of Google and the HITS[2] of Kleinberg are the basic algorithms using link structure information. Generally, these systems function in two steps. In the first stage, a traditional search engine returns a list of pages in response to user query. In the second stage, these systems take account of the links to rank the documents results. In this section we describe some of previous link analysis ranking algorithms.
PageRank (PR), introduced by L. Page and S.Brin [1], which is part of the ranking algorithm used by google precomputes a rank vector that provides a-priori "importance" estimates for all the pages on the Web. This vector is computed once, offline, and is independent of the search query. At the query time, these importance scores are used in conjunction with query-specific IR scores to rank the query results. PageRank simulates a user navigating randomly in the Web who jumps to a random page with probability (1-d) or follows a random hyperlink (on the current page) with probability d. This process can be modelled with a Markov chain, from where the stationary probability of being in each page can be computed.

Intuitively, this formula means that the PR of a page A depends at the same time on the quality and the number of pages which cites A. For example, the pages pointed by the home page Yahoo! that have a higher PR will be judged of good quality. The PR computations are long and require cleaning the entire Web. Moreover, the results obtaining by Google shows that the algorithm witch compute PageRank value of a page is not completely relevant. The query results do not have sometimes any relationship with research carried out. Because search engines does not take into account semantics, context or user profile. From where, the idea to compute personalized PageRank. Last years, research led to three radically different solutions [6], the modular Pagerank, the BlockRank and the Topic sensitive Pagerank. The three approaches approximate PR with some approximation, although they differ substantially in their computational requirements and in the granularity of personalization achieved.

Considering the Web is a nested structure, the Web graph could be partitioned into blocks according to the different level of Web structure, such as page level, directory level, host level and domain level. We call such constructed Web graph as the *block-based* Web graph, which is shown in Fig.2 (left). Furthermore, the hyperlink at the block level could be divided into two types: Intra-hyperlink and Inter-hyperlink, where inter-hyperlink is the hyperlink that links two Web pages over different blocks while intra-hyperlink is the hyperlink that links two Web pages in the same block. As shown in Fig 2, the dash line represents the intra-hyperlink while the bold line represents the inter-hyperlink. There is several analysis on the block based Web graph. Kamvar et al. [18] propose to utilize the block structure to accelerate the computation of PageRank. Further analysis on the Website block could be seen in [13][15]. And the existed methods about PageRank could be considered as the link analysis based on page level in our approach. However, the intra-link and inter-link are not discriminated to be taken as the same weight although several approaches proposed that the intra-hyperlink in a host maybe less useful in computing the PageRank [7].

In [8], Kleinberg introduced a procedure for identifying web pages that are good hubs or good authorities, in response to a given query. To identify good hubs and authorities, Kleinberg's procedure exploits the graph structure of the web. Each web page is a node and a link from page A to page B is represented by a directed edge from node A to node B. When introducing a query, the procedure first constructs a focused subgraph G, and then computes hubs and authorities scores for each node of G (say N nodes in total). In order to quantify the quality of a page as a hub and an authority, Kleinberg associated every page with a hub and an authority weight. Following the mutual reinforcing relationship between hubs and authorities, Kleinberg defined the hub weight to be the sum of the authority weights of the nodes that are pointed to by the hub, and the authority weight to be the sum of the hub weights that point to this authority.

3 Modeling the context of documents

Considering a graph of HTML pages where hypertext links relate source pages to destination pages, and considering the HTML anchor text of a source page that pro-

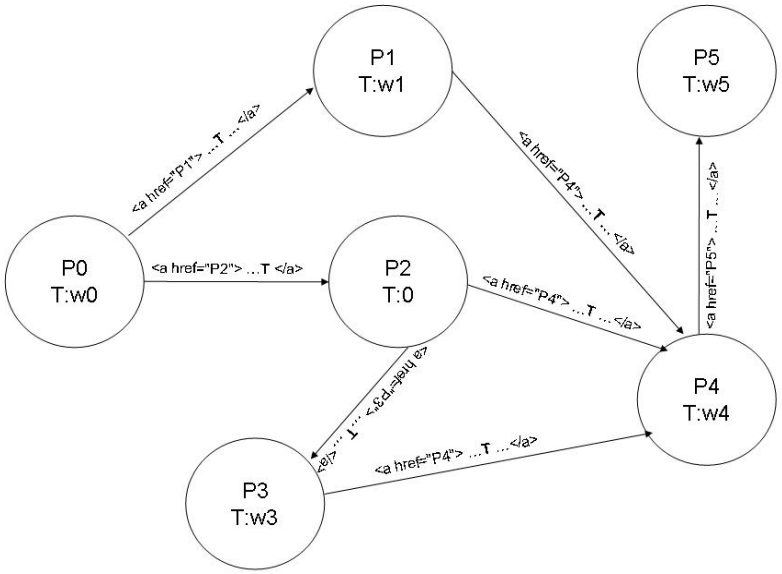
vides information to the destination page. HTML anchors are often surrounded by additionally text that seems to describe the destination page appropriately. The anchor text and the text surrounding an anchor text of a link ("link-context") is used for a variety of tasks associated with Web information retrieval. For example, it may be used by a search engine to rank a page. These tasks can benefit by identifying structural regularities that appear around links and that would constitute a link-context. We describe a framework for conducting such a study. The framework serves as an evaluation platform for comparing various link-context derivation methods. Our focus is on understanding the potential merits of using the zone around the anchor text (link-context), for improving information retrieval. For that, we propose a hyperlink-based term propagation model (HT). The HT model propagates the frequency of query terms in a web page using the context-link information before assigning the relevance weighting algorithms to rank the documents. We consider three types of links: in-link, out-link and in-out-link (bi-directional) (table 1). The HT model can be applied to each type of link by recursively propagating the weight of link-context terms.

Table 1. Applications of the HT model

Weight of link-context	HT propagation function
in-link	$FT^{n+1}(T,D) = FT^{0}(T,D) + \beta * \sum_{(D' \in In(D) \wedge T \in AT(D' \to D)} FT^{n}(T,D')$
out-link	$FT^{n+1}(T,D) = FT^{0}(T,D) + \beta * \sum_{(D' \in Out(D) \wedge T \in AT(D \to D')} FT^{n}(T,D')$
in-out-link	$FT^{n+1}(T,D) = FT^{0}(T,D) + \beta * \sum_{(D' \in In(D) \cup Out(D)) \wedge (T \in AT(D \to D') \vee T \in AT(D' \to D))} FT^{n}(T,D')$

In the Figure 1, we represent an example of a graph of pages where each node represents a page and each oriented arc from node A to node B represents the link-context to B. Each page contains a set of terms whose weight is calculated by combining the Okapi BM25 score and a term weight propagation using the link-context. It is necessary that these terms appear around the anchor text of links between documents. For example, the weight of the term T in the page P4 is calculated from all the weights of the terms of the pages P0, P1, P2 and P3. The strength of each weight depends on the distance between two documents in terms of links. For example, there are three paths between the page 0 and page 5: P0-P1-P4-P5 and P0-P2-P4-P5 of length 3 and P0-P1-P2-P3-P4-P5 of length 4.

Figure 1. Example of link-context

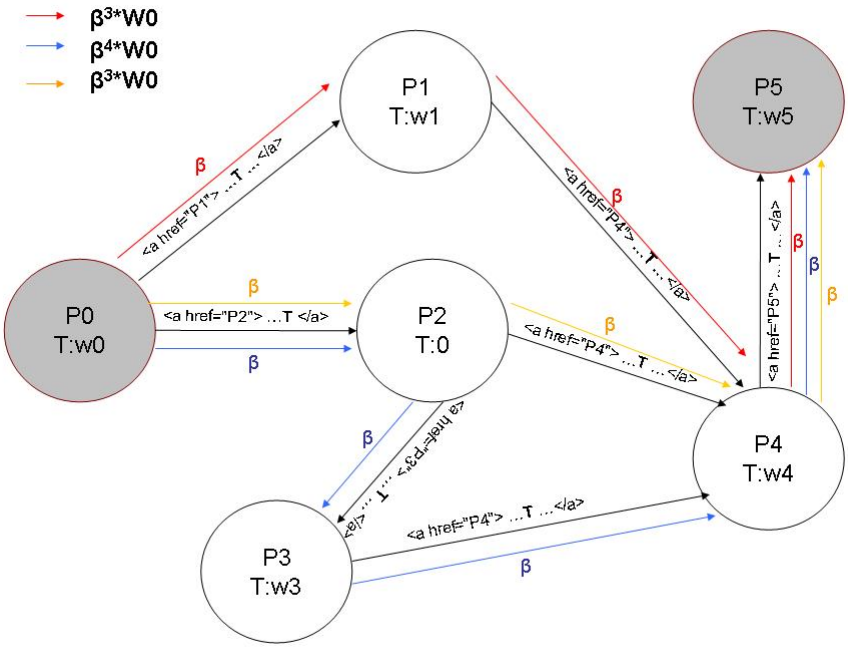


We can easily calculate the weight of the term T in the document D as follow

$$FT^{n+1}(T,D) = FT^{0}(T,D) + \beta * \sum_{(D_i \in In^{-1}(D)} FT^{0}(T,D_i) + \beta^{2} * \sum_{(D_i \in In^{2}(D)} FT^{0}(T,D_i) + ... + \beta^{k} * \sum_{(D_i \in In^{k}(D)} FT^{0}(T,D_i)$$

$In^{k}(D)$ represents a set of documents that are at distance K from document D.

Figure 2. Example of contribution of weight term propagation T from P0 to P5



In table 2, we provide an example of successive iterations corresponding to the figure 1, that illustrates our HT algorithm of weight term propagation. We notice that the propagation weight of terms converge towards the red values. The number of iterations is fixed, in order to eliminate the problem of cycles in the graph.

Table 2. Iterations for the HT model

Iteration 1	**Iteration 2**
$FT^0(P0,T)=W_0$ $FT^0(P1,T)=W_1$ $FT^0(P2,T)=0$ $FT^0(P3,T)=W_3$ $FT^0(P4,T)=W_4$ $FT^0(P5,T)=W_5$	$FT^1(P0,T)=\mathbf{W_0}$ $FT^1(P1,T)= W_1+ \beta* W_0$ $FT^1(P2,T)= \beta* W_0$ $FT^1(P3,T)=W_3$ $FT^1(P4,T)=W_4+ \beta*(W_1+ W_3)$ $FT^1(P5,T)=W_5+ \beta*W_4$
Iteration 3	**Iteration 4**
$FT^2(P0,T)=\mathbf{W_0}$ $FT^2(P1,T)=\mathbf{W_1+ \beta*W_0}$ $FT^2(P2,T)=\mathbf{\beta*W_0}$ $FT^2(P3,T)=W_3+ \beta^2*W_0$ $FT^2(P4,T)=W_4+\beta*(W_1+ W_3)+ 2*\beta^2*W_0$	$FT^3(P0,T)=\mathbf{W_0}$ $FT^3(P1,T)=\mathbf{W_1+\beta*W_0}$ $FT^3(P2,T)=\mathbf{\beta*W_0}$ $FT^3(P3,T)=\mathbf{W_3+\beta^2*W_0}$ $FT^3(P4,T)=W_4+\beta*(W_1+W_3)+ (\beta^3+2*\beta^2)*W_0$

$FT^2(P5,T)=W_5+\beta*W_4+\beta^2*(W_1+W_3)$	$FT^3(P5,T)=W_5+\beta*W_4+\beta^2*(W_1+W_3)+2*\beta^3*W_0$

$FT^4(P0,T)=\mathbf{W_0}$
$FT^4(P1,T)=\mathbf{W_1+\beta*W_0}$
$FT^4(P2,T)=\mathbf{\beta*W_0}$
$FT^4(P3,T)=\mathbf{W_3+\beta^2*W_0}$
$FT^4(P4,T)=\mathbf{W_4+\beta*(W_1+W_3)+(\beta^3+2*\beta^2)*W_0}$
$FT^4(P5,T)=\mathbf{W_5+\beta*W_4+\beta^2*(W_1+W_3)+(\beta^4+2*\beta^3)*W_0}$

4 Experiments over TREC-9

In this section we present an experimental evaluation of our proposed algorithm that we compare to a content based model. We chose the WT10g collection. In our experiments, the precision over the 11 standard recall levels which are 0%, 10%, …, 100% is the main evaluation metric, and we also evaluate the main average precision (MAP) and the precision at 5 and 10 documents retrieval (P@5 & P@10).

Figure 3 shows the experimental results on information retrieval using different context-link methods. The first one which is based on the content-only of the page and is presented with the blue line is the baseline algorithm. The others show results by using our HT model of term propagation according to the types of links. The HT model outperforms the content-only baseline, and specifically the HT model of in-link term propagation is better than the others HT models. These results show that the information conveyed by the in-link is the most important to describe a target page.

Figure 3. Results over TREC-9

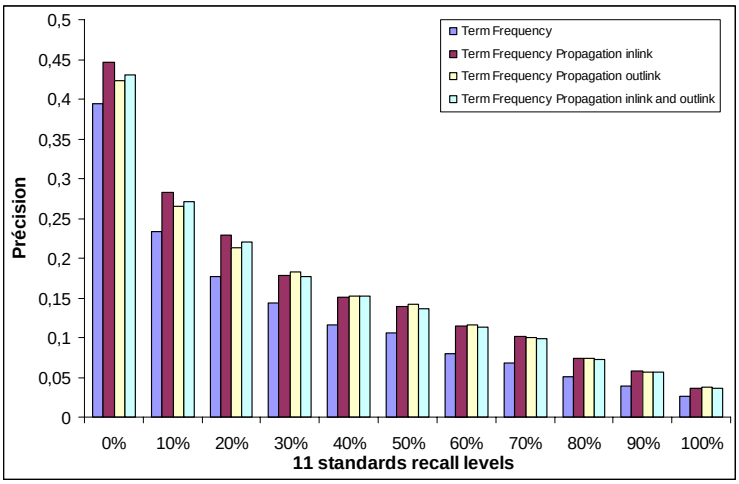


Table 2. Comparisons at MAP, P5 and P10

	TF	TFP_IN	TFP_OUT	TFP_IN_OUT
map	0,1102	0,1416	0,1376	0,1383
P5	0,18	0,22	0,196	0,216
P10	0,148	0,166	0,16	0,16

TF : contents only
TFP_IN : propagation of terms frequency through in-links
TFP_OUT : propagation of terms frequency through.
TFP_IN_OUT : propagation of terms frequency through in-links and out-links.

Table 2 shows that the in-link HT model propagation of terms performs the best result for MAP, P@5 and P@10. For example, the results of in-link HT model propagation achieve 27% for MAP and 22% for P@5.

5 Conclusion

Several algorithms based on link structure to take account of the context of a Web page as an atomic unit of information were developed. But until now, many experiments showed that there is no significant profit compared to the methods based only on content of page. In this paper, we proposed a new method based on link-context using information around the anchor text and the propagation of term weights through the links. We performed experimental evaluations of our system using IR test collection of TREC 9. We conclude that the context of Web pages has a positive impact in the increase of the precision in the top of ranking and in MAP.

We are currently testing our model for expanding queries (relevance feedback) by selecting terms from the surrounding of the anchor text, issued from the co-occurrence matrix between terms of the most relevant documents (we select the top ten relevant documents). Our future work is to test this framework at the semantic blocks level to see the structural effects of blocks on ranking query results. Finally, new measure representing additional semantic information may be explored.

6 References

[1] Brin S. et Page L. (1998), The anatomy of a large-scale hypertextual Web search engine, In Proceeding of WWW7, 1998.

[2] Kleinberg L. (1998), Authoritative sources in a hyperlinked environment, In Proceeding of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[3] Lempel R. et Moran S. (2000), The stochastic approach for link-structure analysis (SALSA) and the TKC effect, In Proceeding of 9th International World Wide Web Conference, 2000.

[4] Savoy J. et Rasolof Y. (2000), Link-Based Retrieval and Distributed Collections, Report of the TREC-9 experiment: Proceedings TREC-9, 2000.
[5] Salton G., Yang C.S. et Yu C.T. (1975), A theory of term importance in automatic text analysis, Journal of the American Society for Information Science and Technology, 1975.
[6] Haveliwala, Taher; Kamvar, Sepandar, Jeh, Glen (2003), An Analytical Comparison of Approaches to Personalizing PageRank, rapport technique, université de Stanford, 2003.
[7] Haveliwala Taher H. (2003), Topic-Sensitive PageRank : A Context-Sensitive Ranking Algorithm for Web Search, Knowledge and Data Engineering, IEEE Transactions on, 2003.
[8] Sepandar D., Kamvar Taher H., Haveliwala Christopher D., Manning Gene H. et Golub (2003), Exploiting the Block Structure of the Web for Computing PageRank, 2003.
[9] Deng Cai; Shipeng Yu; Ji-Rong Wen; Wei-Ying Ma (2004), Block-based Web Search, Microsoft research ASIA, 2004.
[10]Xue-Mei Jiang, Gui-Rong Xue, Wen Guan Song, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma (2004), Exploiting PageRank at Different Block Level - International Conference on Web Information Systems Engineering, 2004.
[11].Jeh G et Widom. J. Scaling personalized web search. In Proceedings of the Twelfth International World Wide Web Conference, 2003.
[12] Porter M.F. (1980), An algorithm for suffix stripping, 1980.
[13] Ji-Rong Wen, Ruihua Song, Deng Cai, Kailhua Zhu, Shipeng Yu, Shaozhi Ye and Wei-Ying Ma (2004), At the web track of TREC 2003, Microsoft research ASIA, 2004.
[14] Doan, B.L. and Brézillon, P. (2004) How the notion of context can be useful to search tools. Proceedings of the World Conference "E-learn 2004", Washington, DC, USA, Nov. 1-5, 2004
[15] Aguiar, F. Improvement of Web Document Retrieval by the Use of Site's Context Hierarchy. In “Intelligent Exploration of the Web". Springer-Verlag, Heildberg, Germany, 2003.
[16] Mark A. Stairmand. Textual Context Analysis for Information Retrieval. Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR '97, Volume 31 Issue SI. ACM Press. July 1997
[17] M. Bazire et P. Brézillon. "Understanding context before to use it". In 5th International and Interdisciplinary Conference on Modeling and Using Context, Lectures Notes in Artificial Intelligence, Vol 3554, pp. 29--40, Springer-Verlag, 2005.