SAAL - A System to Store and Analyze Links

Roberta de Souza Coelho[1], Silvio Romero Lemos Meira[1]

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife, Caixa Postal 7851 – Recife – PE – Brazil

{rsc,srlm}@cin.ufpe.br

Abstract. *The characteristics of the Web environment create many new challenges to Web Information Retrieval Systems - also called search engines (SE). At the same time, unlike "flat" document collections, the WWW is made of hypertexts that provide auxiliary information on Web pages, such as link structure and link text. We have developed a system called SAAL-Sistema para Análise e Armazenamento de Links (A System to Store and Analyze Links) whose main goal is to provide link information to search engines. Such information is obtained by the execution of a link analysis algorithm, called GHHITS (Global Hybrid Induced Hypertext Topic Search) also proposed in this work. Through experiments with real data obtained from a Brazilian SE, we verified that this algorithm can be used to obtain a significant improvement in the retrieval performance of a SE.*

1. Introduction

The World Wide Web is a huge repository of information. The most effective tool to find information on the Web are the Search Engines (SE). To interact with search engines, users submit queries - typically a list of keywords - and receive a list of Web pages ordered by the search engine's internal criteria (ranking function). Recent studies [Davison, 2001] report that the performance of actual search engines is far from being fully satisfactory. While technological progress has made it possible to deal reasonably with the size of the Web in terms of number of indexed pages, the big problem is just to give users the information they need.

The main reason to explain such difficulty may lay on traditional Information Retrieval (IR) techniques. Traditional IR algorithms were developed to deal with relatively small and coherent textual collections such as newspaper articles or book catalogs. The Web, on the other hand, is massive, much less coherent, with great variation in quality and type of information present in its pages. Many pages that contain the search terms may be of poor quality, or not relevant due to spamming [Davison, 2001] techniques. Thus, as Web pages are not sufficiently self-descriptive, techniques that base their decisions just on the pages' content are easy to manipulate. Even the most sophisticated textual techniques suffer from an intrinsic weakness: they do not take into account the Web structure the page is part of. This requires new IR techniques, or extensions to the old ones, to deal with Web data.

As in citation analysis of published works, the most influential ("important") documents on the Web have many other documents recommending (linking) them. Some new algorithms have been proposed to explore Web link structure. IBM's

CLEVER project [Chakrabarti et al., 1998], Stanford's Google [Brin and Page, 1998], and the Web Archaeology research [Bharat and Henzinger, 1998] at Compaq's Systems Research Center have demonstrated some of the contributions that Web link analysis can bring to search engines. The information discovered by these algorithms have been used as new criteria to rank Web pages. Therefore, our thesis concentrated in the important issue of Web link analysis, which can bring several benefits to Web information retrieval.

In this context, we propose a link analysis algorithm, called GHHITS (Global Hybrid Induced HITS), based in those previously cited algorithms. To validate this algorithm we built a system, called SAAL - Sistema para Análise e Armazenamento de Links - to store Web links and execute GHHITS link analysis algorithm on them. Through experiments with real data obtained from a Brazilian SE, we verified that this algorithm can be used to improve the retrieval performance of a SE.

This paper is divided as follows. Section 2 presents a brief overview to Indexer-Crawler Centralized Architecture as a background to introduce the proposed algorithm in Section 3. Section 4 describes the experiments made to validate the algorithm. And discusses the experimental findings. Finally, Section 5 concludes this paper by discussing the contributions and future work to be developed from the work started in our thesis.

2. Indexer-Crawler Centralized Architecture

Most search engines use an Indexer-Crawler centralized architecture (explained in [Baeza-Yates and Ribeiro-Neto, 1999]). A simplified diagram of it is shown in Fig. 1 (a). The Crawler-Indexer module consists of the Crawler, the Indexing Server and the Link Server, and is responsible for collecting and storing Web pages. The Search module consists of the User Interface, the Query Engine and the Ranking function.

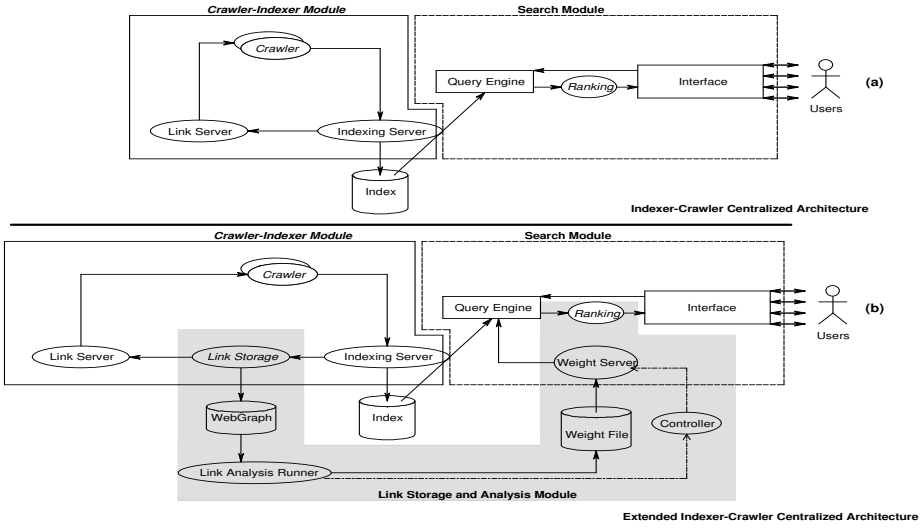

Figure 1: The Indexer-Crawler Centralized Architecture and an extended version using SAAL.

A problem faced by this architecture is that it discards the Web link structure during indexing and search. To address this problem we built a system, called SAAL,

which extends the Indexer-Crawler centralized architecture, with 6 new elements grouped in a Link Storage and Analysis Module, shown in Fig.1(b) and explained below.

The Link Storage receives link information and applies heuristics to eliminate nepotistic links [Davison, 2001] . The Web Graph [Coelho, 2002] stores the link structure. The GHHITS Runner applies GHHITS algorithm (detailed in Section 3) in the Web Graph. This algorithm assigns a "importance" weight to each page stored in the Web Graph. The Weight Server is then responsible for responding weight requests. The Controller is responsible for starting, stopping e reloading the Weight Server. The new Ranking function takes into account the "importance" weight together with "textual" weight, generated by conventional IR algorithms. In the formula below the "importance" ranking is represented by the variable AUT, the "textual" ranking is represented by the variable TEXT - both normalized by the highest weight - a and b are constants empirically determined during the experiments detailed in Section 4.

$$Ranking = (a * TEXT + b*AUT) \quad (1)$$

3. Proposed Algorithm

The algorithm proposed in this work is based on previously cited researches: IBM's CLEVER project, Stanford's Google, and the Web Archaeology research at Compaq's Systems Research Center. Some advantages and disadvantages of each research is discussed below.

PageRank, the link analysis algorithm used by Google, is based upon the notion that important pages are those that receive links from other important pages. This algorithm is global, and it is applied off-line over the whole Web graph. The major advantage of this approach is that there is no additional run-time link analysis penalty during a query search process. However, one disadvantage is that, in many cases, the relationship that exists among sites suggests that a different model is at work: when two sites competes do not reference each other.

Web Archaeology research and the CLEVER project are extensions of HITS (Hyperlink Induced Topic Search) algorithm proposed by [Kleinberg,1998]. HITS family algorithms recognizes the "importance" of competing pages, based on the existence of pages that act as resource lists ("bookmarks") linking to all competing sites from a single point. This algorithm is based on two concepts: (i) the authority pages - pages that are highly referenced and are those most likely to be relevant to a particular query; (ii) the hub pages - pages that are not necessarily authorities themselves but point to several authority pages. HITS associates two weights to each page: authority weight and hub weight. It is executed on a subset of pages originated from a specific query, and needs to be computed on-line every time a query is submitted to the SE.

In our work, we developed GHHITS (Global Hybrid HITS). It is a query independent (global) link analysis algorithm applied off-line over the whole indexed Web Graph. GHHITS associates an "authoritative" and a "hub" weight to each page of Web Graph and applies some heuristics proposed by the previously cited algorithms. The formulas below represent the idea of GHHITS algorithm.

$$a(i) = InitAut(i) + \sum_{i \in B(i)} h(j) * aut\_wt(j,i) \quad \textbf{(2)}$$

$$h(i) = InitHub(i) + \sum_{i \in F(i)} a(j) * hub\_wt(i,j) \quad \textbf{(3)}$$

Where:

- (i,j) represents a link from page i to page j;
- B(i) and F(i) are sets of all backlinks and forwardlinks of page i, respectively;
- aut_wt e hub_wt are fractional weights assigned to links to make documents on a single host have the same influence as a single document would on the computation of hubs and authorities. Those weights were proposed by Compaq's project;
- InitAut e InitHub are personalization vectors used to assign a higher weight to pages previously known as "authoritative" pages and "hub" pages according to a Web user or a Web community. A similar idea is adopted by PageRank.

4. Experiments Results

To verify the impact in retrieval performance of using the "authority" weight in Ranking function, we perform experiments using SAAL integrated with Radix (www.radix.com), a search engine covering the Brazilian Web. During this experiments 3.742.983 pages were indexed - (11,14% of Brazilian Web in January/2002 according to statistics defined by [Alonso et al., 200]). The WebGraph was composed by 2.882.853 links which consisted of indexed pages and its forwardlinks which were not considered nepotistic links [Davison, 2001]. GHHITS algorithm was executed in this subset, the list of most authoritative pages were from de domain ".com", this characteristic may be explained as a side effect of globalization - http://www.cin.ufpe.br /~rsc/HITS/Top100.html. Our experiments consisted of evaluating 5 different combinations of the "importance" and "textual" weights described in formula (1). Each combination assigned different values to a and b. Such combinations were the following: TEXT; AUT; 0.85 x TEXT + 0.15 x AUT; 0.75 x TEXT + 0.25 x AUT; 0.65 x TEXT + 0.35 x AUT. To evaluate the retrieval performance of these ranking functions we defined an evaluation methodology described in [Coelho, 2002]. The metrics chosen by this methodology was: TREC Style Average Precision at cutoff 10 (TSAP@10) and relative precision at cutoff 10 (p@10). Tables 1 and 2 show the metrics values obtained by each strategy during the experiments. They also show the values of relative comparison (RC-TEXT) and absolute comparison (AC_TEXT) in relation to values obtained by TEXT.

Table 1. Values of TSAP@10 and p@10.

Estratégia	TSAP@10	AC-TEXT	RC-TEXT	p@10	AC-TEXT	RC-TEXT
TEXT	0,332	-	-	0,478	-	-
AUT	0,166	-16,6%	-50,05%	0,284	-19,4%	- 40,52%
0.85TEXT_0.15AUT	0,368	3,6%	10,74%	0,516	3,9%	8,13%
0.75TEXT_0.25AUT	0,381	4,9%	14,64%	0,523	4,5%	9,42%
0.65TEXT_0.35AUT	0,366	3,4%	10,13%	0,509	3,2%	6,65%

These experiments showed that all strategies which used two sources of information, "importance" weight in conjunction with "textual" weight,

(0.85TEXT_0.15AUT, 0.75TEXT_0.25AUT, 0.65TEXT_0.35AUT) obtained a better retrieval performance, according to precision@10 and TSAP@10, then the "textual" only strategy (TEXT). Those three strategies did not assign a value to b higher than 50% (b>0,5), because we aimed to use link information as a auxiliary source of information, not as the majority one - the one that influences mostly in the Ranking weight. The strategy which used only the "importance" weight (b=1, a=0) to rank Web pages obtained the worst performance among all experiments. It was already predictable, once AUT strategy discards important characteristics of a page like: tfidf [Baeza-Yates and Ribeiro-Neto, 1999],distance between terms and term positions.

The strategy 0.75TEXT_0.25AUT showed the most significant improvement in retrieval performance in relation to TEXT strategy, 9,42% and 14,64% according to precision@10 and TSAP@10 respectively. This strategy represents the equilibrium between this two sources of information, the "importance" weight and the "textual", weight in the evaluated scenario. The strategies which assigned a lower and a higher value to variable b (0.85TEXT_0.15AUT, 0.65TEXT_0.35AUT) achieved a lower improvement in retrieval performance.

5. Concluding Remarks

We obtained significant improvement in retrieval performance in all strategies that used the “importance” weight in conjunction to “textual” weigh. Such results demonstrated that GHHITS, can be used in conjunction with a textual strategy to improve the search engine retrieval performance. This is a motivation to improve this work: (i) use the link text together with the link structure during the calculation of hub e authority weights; (ii) discover the known important pages and use this information to refine the values assigned to initAut and initHub vectors which in this first experiments the assigned the same values to all pages present in the subset - Radix uses the known important sites when updating indexed sites [Barbosa et al, 2002].

References

Alonso, E., Silva, A., Golgher, P., Barra, R., Laender, A., Ribeiro-Neto, B., Ziviani, N.(2000).Um Retrato da Web Brasileira. SEMISH'2000.

Baeza-Yates, R., Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.

Bharat, K., Henzinger, M. (1998). Improved Algorithms for Topic Distillation in Hyperlinked Environments. ACM SIGIR, pages 104-111.

Brin, S., Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of 8th WWW Conference, pages 107-117.

Chakrabarti, S., Gibson, D., and Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. www7th.

Coelho, R., Fonseca, M., Meira, S.(2002). Combinando Informações Textuais e Estruturais na Recuperação de Documentos Web. RITA (Vol. IX , number 3).

Davison, B. (2001) Recognizing nepotistic links on the Web. Artificial Intelligence for Web Search, Technical Report WS-00-01, pages. 23-28, AAAI Press.

Kleinberg, J.,(1998).Authoritative Sources in a Hyperlinked Environment. Proceedings of the 9th ACM-SIAM, pages 668-677.

Barbosa, L. A., Ramalho, F. S, and Salgado, A. C. (2002) "Atualizando Informação Dinâmica na Web: o Caso do Conteúdo de Notícias". RITA (Vol. IX , number 3).