Improving Web Site's Accessibility

Garofalakis John
University of Patras, CTI
GREECE
E-mail: garofala@cti.gr

Kappos Panagiotis
University of Patras, CTI
GREECE
E-mail: kappos@cti.gr

Makris Christos
University of Patras, CTI
GREECE
E-mail: makri@ceid.upatras.gr

ABSTRACT

We consider the problem of improving the performance of web access by proposing a reconstruction of the internal link structure of a web site in order to match the quality of the pages (measured in terms of their link importance in the web space) with the popularity of the pages (measured in terms of their importance recognized by web users). We provide a set of simple algorithms in order to increase the access rate of popular pages by using local reorganization of the web site's pages.

Keywords

Hypertext Linking, Web Performance, Search Engines.

1. INTRODUCTION

Search engines index web documents using their own proprietary techniques for web information retrieval. We can summarize the most common functionalities of modern search engines [5,6]: (i) they use a spider or crawler that traverses the web for fetching new pages to be included into the repository, or just for updating existing data, (ii) they use an indexer that helps them do full text indexing of web pages, (iii) they provide the user with the ability to add new URLs for exploration, (iv) they use a ranking mechanism, based on the indexer, for sorting the produced results of a search query.

Recently there have been proposed search engines or research prototypes (see [1,2,7,8,9]), that follow other criteria on match ordering. These search engines try to exploit the link structure of the web, in order to uncover the importance of a page for a given query topic. The ranking is based on the link interconnection of the web space, which is a significant improvement towards defeating advertising "spammers". In [2] (see also [8] for a related approach) the link importance of a web page is equal to the sum of its in- and out-degree. In [9] the rank of a page p is computed by a method that incorporates the ranking of pages reachable from p. The Google search engine uses a calculated number called PageRank [1] that is indicative to a web page, in accordance to the whole number of existing web documents. Finally in [7] the link importance of a page is computed by using a two-way feedback process that distinguishes two kinds of pages: hubs (pages with high quality outgoing links) and authorities (pages with high quality incoming links).

It is apparent that there is a difference between the *"real web space"* of information and the space defined by the search engines, when it is to say about specific thematic category information. An obvious reason is that not all web sites are contained in search engines and not all links are always ranked properly. The main reason, though, is that the link importance of a web page is *different* from the importance of that particular page, as far as the users on the web are concerned. On the contrary, web site's log files represent the *real* situation of the user's acceptance, for a specific web page or site. Link matches in search engines specify a different information space. A lot of surveys have covered many aspects of user habits, ranging from age to social status of web users and from political preferences to security issues. However, only little work, based on log file analysis data of a web site, has been conducted to unveil the most favored kind of resources the users prefer to access. The development, maintenance and analysis of web site's log-files enlighten many useful aspects of the way users traverse the web and click through links. The log files of a web site directly reflect the *"real world"* acceptance of the services, provided by a web resource. They show the true impact towards the potential user querying the specific resource.

In our situation we have a specific web site that currently serves a number of hits or pages accesses. It is better to talk about page accesses, since that metric is much more objective than the hits parameter, which counts any type of file resource including user interface pictures etc. By analyzing the web site's log files, the most popular pages can be found. We want to maximize the page accesses to our web site, meaning that we want to increase the traffic and potentially the importance of our site's content for the benefit of the users. In this paper, we discuss the methods to achieve this by restructuring the web site's link space to favor popular pages (pages with large number of accesses), so that pages with link importance can transfer an amount of their value to pages, *in the same site*, that are more popular (from a casual user perspective). The whole process can be rephrased as an attempt to balance the two metrics, for each page in our site, since imbalance in the two metrics is a sign of bad web site design. Our effort can be considered as an attempt to use information gained from traffic patterns to increase the accessibility of the web space.

To achieve this target, we need to sketch methods and algorithms that will be as general as possible. In other words, to be valid for any kind of search engine independently of the kind of link metric it uses. Our approach is based on the observation that the link importance of a web page is reflected to the link quality of its incoming and its outgoing links, so copying or transferring these links to other pages is a meaningful way to increase the link importance value of these pages.

In section 2 we define the notion of page popularity, in section 3 we describe three algorithms that can be used to improve the accesses of a web site for a search engine that is based on link metrics. Section 4 shortly presents experimental results.

2. PAGE POPULARITY

One easy way to estimate a web page's popularity is to count the accesses to this page based exclusively on a given log file. However, counting these *absolute accesses (AA)* from the log file may be misleading. The factors that must be taken into consideration when popularity is to be counted are: (i) the depth of the page (how many steps it is from the home page), d, (ii) the number of pages at the same depth as the page being examined, n_d, (iii) the number of references (hyperlinks) to this particular page from other pages of the site, r. Let's assume that there is a factor a that embraces all the above parameters. In [3]

a new term called *relative accesses (RA)* was introduced, which was derived from the following equation: $RA = a * AA$.

The relation between a_i, d_i, n_i and r_i , the corresponding values of a, d, n and r for page i, should have the following form:

$a_i = F(d_i, n_i, 1/r_i)$, $i = 1,\ldots,K$, K is the number of web pages

One possible way of specifying the coefficient $\boldsymbol{a_i}$ can be the following equation: $\boldsymbol{a_i = c_1 * d_i + c_2 * n_i / r_i}$

The definition of parameters $\boldsymbol{c_1}$, $\boldsymbol{c_2}$ is definitely a challenging issue. Currently, we are in the position to know that $\boldsymbol{c_1}$, $\boldsymbol{c_2}$ can affect significantly algorithm's behavior. The higher the value of $\boldsymbol{c_1}$ is, the more vigorously pages deeper in the HTML structure are promoted. This can lead sometimes in cases (when value of $\boldsymbol{c_1}$ is extremely high), where the arrangement of pages is "volatile", and after each link-editing algorithm execution, a new re-arrangement occurs. Definitely, this is not a sound case, since any robust algorithm should achieve a "balanced" page-arrangement (where any further execution of the algorithm does not affect the structure) in one or a few executions. If we assign the value 1 to parameters $\boldsymbol{c_1}$, $\boldsymbol{c_2}$ the above equation can be rewritten as below: $\boldsymbol{a_i = d_i + n_i / r}$

3. IMPROVING ACCESSIBILITY

The first algorithm tries to improve the link importance of popular pages for the case where the link importance depends strongly on the *incoming* links to the page, the third algorithm tries to increase the link importance of popular pages, for the case where the link importance depends strongly on the *outgoing* links of the page and the second algorithm works independently regardless the used link metric.

Algorithm 1

The idea behind the algorithm is to put additional links to the pages of the site with high link importance that point to the most popular pages. It consists of the following phases: (1) Calculate the link importance of the pages in the web site (by using for example the algorithms in [1,2,8] or the algorithm in [7] for selecting good authorities) and select a subset S of these pages with the higher link importance, (2) Calculate the page popularity of all web pages in the site, (3) Select the most popular pages, (4) Create hyperlinks from the pages selected in phase 1 to the most popular pages.

Algorithm 2

This algorithm is based on the swapping of the pages, in the web site hierarchy, by interchanging the file names of popular and link important pages (with similar content). It consists of five phases: (1) Calculate the link importance of the pages in the web site (we can use any of the already proposed link metrics) and select a subset of these pages with the highest rank, (2) Calculate the page popularity of all web pages, (3) Find all external links pointing to all of our web site's pages, (4) Select the most popular pages, (5) Interchange the file names of the popular pages with a subset of the highest linked ranked pages and of the pages with the biggest number of external links.

Algorithm 3

This algorithm tries to improve the link importance of popular pages, for the case where the link importance depends strongly on the outgoing links of the page. It consists of the following four phases: (1) Calculate the link importance of the pages in the web site (by using for example the algorithm in [9], or the algorithm in [7] for finding good hubs), (2) Calculate the page popularity of all web pages, (3) Select a set of links out of the highest ranked pages from phase 1, (4) Copy the links selected in the third phase, to the most popular pages.

4. EXPERIMENTAL RESULTS

For the verification and validity of the three algorithms we applied them to a real working commercial web site. This web site operates at the http://www.jokes.gr address and is one of the top 7 most highly accessible commercial content web sites in Greece serving more than five million page accesses per month. The total duration of the measurements was 68 days. The duration of the first (before the link changes) and the second (after the link changes) measurements was 34 days and can be considered a representative amount of time to show the real user preferences, for the type of material of the web site, taking in mind that approximately five million of web pages were transferred to 130.000 unique IPs, having 15 pages as the mean user session length. The results of the experiments (for more details see [4]) were quite promising since our algorithms increased the accessibility of already popular pages. We enlarge our test-bed database constantly and intend to present the results of our work shortly. There are a lot of open research areas to be carried out since web site reorganization and linking techniques should be much more deployed. The addition of new parameters and metrics that could affect the internal web site design is also within our future work.

5. REFERENCES

[1] S. Brin, L. Page, *"The anatomy of a large-scale hypertextual web search engine"*, Proceedings of the 7th International World Wide Web Conference, April 1998.

[2] J. Carriere, R. Kazman, "*WebQuery: Searching and visualizing the Web through connectivity*", Proceedings of the 6th International World Wide Web Conference, 1997.

[3] J. Garofalakis, P. Kappos, D. Mourloukos, "*Web Site Optimization Using Page Popularity*", IEEE Internet Computing, July-August 1999, pp. 22-29.

[4] J. Garofalakis, P. Kappos, C. Makris, "*Improving the Performance of Web Access by Bridging Global Ranking with Local Page Popularity Metrics*", CTI T.R. 2001/01/03.

[5] Gudivada, Raghavan, Grosky, Kasanagottu, "*Information Retrieval on the World Wide Web*", IEEE Internet Computing, September-October 1997, pp. 58-68.

[6] M. Henzinger, "*Web Information Retrieval – an Algorithmic Perspective*", Proceedings of the 8th Annual European Symposium on Algorithms, (ESA) September 2000, pp.1-9.

[7] J. Kleinberg, *"Authoritative sources in a hyperlinked environment",* Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. Also appeared as IBM Research Report RJ 10076, May. 1997.

[8] Y. Li, "*Towards a qualitative search engine*", IEEE Internet Computing, 2(4): 24-29, 1998.

[9] M. Marchiori, "*The quest for correct information on the Web: Hyper search engines*", Proceedings of the 6th International World Wide Web Conference 1997.