Journal of Web Engineering, Vol. 2, No.1&2 (2003) 090-104
© Rinton Press

TECHNIQUES AND METRICS FOR IMPROVING WEBSITE STRUCTURE[a]

ELENI CHRISTOPOULOU[1,2], JOHN GAROFALAKIS[1,2], CHRISTOS MAKRIS[1,2], YANNIS PANAGIS[1,2], ATHANASIOS PSARAS-CHATZIGEORGIOU[1], EVANGELOS SAKKOPOULOS[1,2] AND ATHANASIOS TSAKALIDIS[1,2]

[1]Department of Computer Engineering & Informatics
School of Engineering, University of Patras
Rio Campus, 26500 Patras, Greece
[2]Research Academic Computer Technology Institute
Internet and Multimedia Technologies Research Unit
61 Riga Feraiou Str. 26110 Patras, Greece

{hristope, garofala, makri, panagis, psaras, sakkopul, tsak}@ceid.upatras.gr

Received April 16, 2003
Revised May 28, 2003

Evaluation of the link structure of a web site and its redefinition to achieve increased efficiency with regard to easier information retrieval is a common problem in website development. Nevertheless much effort has been devoted in order to analyze the overall statistical properties of a web site, rather than to assess the actual value of its pages. In this paper two distinct metrics are proposed, which aim to quantify the importance of a web page based on the visits it receives by the users and its location within the website. Subsequently, certain guidelines are presented, which can be used to reorganize the website, taking into account the optimization of these metrics. Finally we evaluate the proposed algorithms using real-world website data and verify that they exhibit more elaborate behavior than a related simpler technique.

Key words**:** Web Metrics, Web Organization, Log File Processing

Communicated by: M Gaedke & G. Rossi

1 Introduction

The free-for-all Internet has led towards a massive publication of information. The Web nowadays seems to serve as a worldwide free library. This chaotic nature of the WWW is mirrored in the structure of websites, making it essentially haphazard.

The ad hoc design, characterizing most sites, can presumably lead to the existence of "concealed" information, either due to lack of foresight or due to incorrect assessment of the presented information importance. As a consequence such website structures are usually neither user friendly nor easily maintained. To address these problems, website designers would appreciate a method that can guide them to create a structure, which is user friendly and without "concealed" information, and, complementary, a method to assess an existing site with regard to these aspects.

This paper focuses on evaluating methods of already existing websites, based on user behavior. We extend and add on the fundamental ideas outlined in [9]. According to that method, the number of hits a page receives, as those are calculated from log file processing (see [7] for an early analysis on

[a] Preliminary work has appeared as poster presentation in the Twelfth International World Wide Web Conference, WWW2003, 20-24/5/2003, Budapest.

web logs), is not a reliable metric to estimate the page's popularity. Thus a refined metric is proposed, which takes into account structural information. Based on this new notion of popularity, *reorganization* of certain pages is proposed. After performing some reorganization steps, they also observed an overall improvement to site access.

Acknowledging the importance of reorganization proposals, our aim is to further facilitate the idea of reorganization by introducing two new metrics. The first ones takes into account both structural information and the differentiation between users coming from within the website and users coming from other websites, while the second uses a probability model to compute a suitable refining factor. Key feature of the new metrics is the higher fidelity on the proposed reorganization proposals. We evaluate and examine these metrics by comparing them with the previous one.

The presentation of techniques and metrics for website optimization is organized as follows: Discussion of tools, applications and previous scientific work is included in section 2. In section 3 we discuss the proposed ways to estimate page popularity. In the sequel in section 4 we describe the rules and methods that were used to produce actual-experimental proposal lists. Section 5 provides a comparative assessment of the results indicating in this manner different possibilities in web site reorganization treatment. Conclusions and future directions are mentioned in section 6.

2 Related Work

User visits' analysis is the first step in any kind of web site evaluation procedure either when it involves re-design and reorganization or not. There are in fact a large number of commercially available systems that analyze a web site's traffic. Unfortunately the majority of these tools are limiting their actions to statistical reports (e.g. Analog [16], Weblogs [19], Web Trends [18], SurfStats Log analyzer [17] etc). All of them provide the web administrator with a set of statistics and several details about the most visited files and pages (by absolute access numbers). Our intention is to make a step forward beyond this kind of simple reports and to provide higher fidelity in the proposed results.

Even though researchers have taken interest in the problem fairly recently, there has also been significant scientific work in the field of web site reorganization. Some early work has been presented by Chen et al [5]. It focuses on the extraction of user behavioral patterns from log files. Work on user path analysis in a web site has been done by Berkhin et al [1]. Their goal was to understand visitors' navigation within the site. Later, Srikant et al [14] proposed an algorithm, which aimed to solve the problem by automatically discovering all pages in a website whose location is different from the location where visitors expect to find them. Several years ago, structural analysis of the hypertext graph and corresponding metrics has been presented by Botafogo et al [3]. They have presented several metrics, i.e. measures to evaluate a hyperstructure. Recent work on the hypertext graph, focuses on websites; Zhou et. al. [15], use graph models to represent the site structure. Weights on these graphs are used to evaluate and improve website structure. Bose et. al. [2] introduced the concept of *HotLinks*, links added to the site structure in order to minimize the expected number of steps, required to access a certain page. They postulate however, that pages of interest are only the leaves of the tree and that the access distributions are known.

On the contrary, little has been done on implementing the proposals that are listed by the algorithms as the ones mentioned above. Only some ad-hoc tools exist, implemented by the corresponding authors in order to perform some reorganization for experimental purposes. Evidence of a semi-automatic reorganization framework (approval by the web designer is still needed) is given in

[6]. In this current work we analyze the proposal lists' production conditions and the different ways that proposed reorganization may be implemented.

3 Estimating Page Popularity

Given the fact that we can count the *Absolute Accesses* (AA_i) to a specific page *i* of a site, it is a challenge to use this information in order to provide a more objective measure of how important this page is. Garofalakis et. al. [9] defined this measure as *Relative Access* (RA_i) and stated that:

$$RA_i = a_i * AA_i \tag{1}$$

That is, the *RA* of page *i* is a result of the multiplication of AA_i by a coefficient a_i. The purpose of a_i is to skew AA_i in a way that better indicates a page's actual importance. Hence, a_i incorporates topological information, namely page depth d_i within site, the number of pages at the same depth n_i and r_i, the number of pages within site pointing to it. Thus $a_i = d_i + n_i/r_i$.

Yet effective, this approach is rather simplistic. We follow the concept of Relative Accesses, but we give in this section two refinements of the original approach.

3.1 Redefining a_i

It is crucial to observe that a web page is accessed in four different ways. First it gets accesses within site, second directly via bookmarks, thirdly by incoming links from the outside world and finally by typing directly, its URL. Under this observation, we can decompose a_i, into two factors, $a_{i,in}$ and $a_{i,out}$. The first one, $a_{i,in}$, reflects the ease of discovering page *i*, under the specific site organization. The harder it is to end up to page *i,* when browsing the site, the higher we want $a_{i,in}$ to be. Thereby, $a_{i,in}$ boosts *i*'s absolute access figures, indicating that page *i* was more important for its visitors than other, easier to discover pages. This factor is similar to [9], but with newly introduced composition.

The new factor that we introduce, $a_{i,out}$, designates the importance of a specific page for the outside world, i.e. pages that point to *i* from other domains and bookmarks to *i*. The more links from other domains or bookmarks a page gets, the higher value $a_{i,out}$ shall be assigned, in ways that we will study. We feel that accesses from the outside world are of great importance to a specific page, giving an objective measure of the page's actual importance. Therefore, it is essential to introduce $a_{i,out}$. In what follows we discuss two approaches in defining $a_{i,in}$ and $a_{i,out}$.

3.2 Using Topological Information

Suppose we have a site with *s* pages, organized in a tree as shown in *Figure 1*. We define the following quantities. Let d_i be the tree-depth of the page *i* and d_{max} the maximum tree depth. Let n_i denote the outdegree of page *i*, i.e. the number of links leaving it and n_{max} the maximum outdegree for a page of the site. We say that page *j* is the parent of page *i* and we denote it by *j=p(i)*, if there is a link from page *j* to page *i*. This relation shall follow from the initial hierarchical organization of the site. The problem is that e.g. page *i* could be pointed from more than one pages within site. As parent, we denote the page resulting in, say, a DFS (Depth First Search) fashion, and the other r_i links are just crossing edges. Under these assumptions $a_{i,in}$ is defined as:

$$a_{i,in} = \frac{d_i}{d_{\max}} + \frac{n_j}{n_{\max}} + \frac{1}{r_i + 1}, j = p(i) \tag{2}$$

Thus, $a_{i,in}$ depends on the depth of page *i* (the deeper *i* lies the more its discovery gets value), the number of *i'*s siblings (many siblings raise the importance of the *i*'s choice) and the number of other pages pointing to it (few ways to access *i*, so the access figures it receives shall be raised). We can observe from Equation 2 that, since we normalize each influence factor, $a_{i,in} \leq 3$.

We need some extra definitions, in order to deliver a formula for $a_{i,out}$. Let B_i, be the number of bookmarks for page *i*, and B_{max} the maximum number of bookmarks for a site page. Furthermore, let L_i be the number of links from other web pages to page *i* and $L_{max} = \max_{i \in \{1,\ldots,s\}} L_i$. We define $a_{i,out}$ as,

$$a_{i,out} = \frac{B_i}{B_{max}+1} \cdot \frac{d_i}{d_{max}} + \frac{L_i}{L_{max}+1} + 1, \tag{3}$$

with d_i, d_{max} as previously defined. Equation 3 implies that $a_{i,out}$ depends on the number of bookmarks for page *i*. When a page has many bookmarks this indicates high popularity among the visitors of our site, which is exactly what we want to capture. Moreover we include the factor d_i/d_{max}, giving a weight to the number of bookmarks w.r.t. the page's depth. Our motivation for that is that for instance, the root page might get a lot of bookmarks but we already expect this page to be popular. On the other hand when a page some levels down the site hierarchy has bookmarks, this page might be important. We use the same idea regarding the links from outside. When a page is pointed to by many independent sites, it might be objectively important. The constant additive factor at the end, comes in order to make $a_{i,out} \leq 3$, just as with $a_{i,in}$, so that they both have the same maximum potential influence on RA_i (see Equation 4).

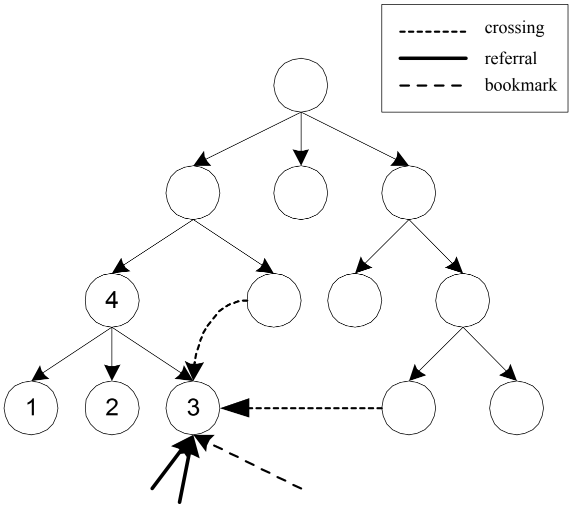

Figure 1. The site's tree infrastructure

In the example organization shown in *Figure 1*, we have $d_3 = d_{max} = 4$, $n_{p(3)} = n_{max} = 3$ and $r_3 = 2$, so $a_{3,in} = 7/3$. We also have $B_3 = 1$, $L_3 = 2$ and say, $B_{max} = 4$, $L_3 = 3$ respectively. Thus, $a_{3,out}$ = 1/4*1+2/3+1 = 23/12.

In order to apply our newly derived factors, we also separate AA_i, into two parts: $AA'_{i,in}$ and $AA_{i,out}$, that keep the accesses from inside the site, *exclusively* for page *i* (and not for its children) and the accesses having an antecedent from another domain, respectively. Finally, the relative access for page *i, RA_i* is:

$$RA_i = a_{i,in} \cdot AA'_{i,in} + a_{i,out} \cdot AA_{i,out} \tag{4}$$

3.3 Using Routing Probabilities

We are about to follow now, another approach for computing a_i's constituents. Our focus is primarily on defining the meaning of $a_{i,in}$, making therefore some simplifying assumptions about other type of accesses.

It is tempting to model traffic inside a site, using a *random walk* approach (e.g. like those in [12],[3]) but $a_{i,in}$, as defined in Section 3.1, models the ease (or difficulty) of access that a certain site infrastructure imposes to the user. Thus, in this section we introduce access probabilities. The idea is to increase a page's relative weight, inversely proportional to its access probability.

Firstly, we introduce some terms. The site structure can be thought as a *directed acyclic graph* (DAG) *G*, as shown in *Figure 1*, with *s* nodes and v_i denoting page *i*. Suppose that a user starts at the root page v_r, looking for an arbitrary site page v_t. At each node v_i he makes two kinds of decisions: either he stops browsing or he follows one of the $out(v_i)$ links to pages on the same site. If we consider each kind equally probable, the probability p_i, of each decision is $p_i = (out(v_i)+1)^{-1}$.

Consider a path $W_j = \{v_r, v_1, \ldots, v_t\}$, from v_r to v_t. Counting the routing probabilities at each step, the probability of ending up to v_t via W_j, is simply:

$$P_{t,j} = \prod_{\forall i, v_i \in W} p_i$$

There may be more than one paths leading to *t*, namely $W_1, W_2, \ldots, W_k$. The overall probability of discovering *t*, D_t is:

$$D_t = \sum_{i=1}^{k} P_{t,i} \tag{6}$$

For the example of *Figure 1*,

$$D_3 = \frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{4}+\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}$$

Considering page *i* as target, the highest D_i is the lower $a_{i,in}$ shall be, so we choose $a_{i,in}$ to be, $a_{i,in} = 1 - D_i$. We also let $a_{i,out}$ to be one. Thus we define RA_i as:

$$RA_i = (1 - D_i) \cdot AA'_{i,in} + AA_{i,out} \tag{7}$$

with $AA'_{i,in}$ and $AA_{i,out}$ defined as previously.

Observe that by letting $a_{i,out}$ =1 we achieve the same effect as with counting each external access, as a path of length one. Though, the aggregate of all individual access probabilities is no more a probability. Note also that the access paths notion embodies the d_i and r_i factors of the previous section as the number of factors in (5) and (6) respectively. Furthermore, it always holds that $a_{i,in}$<1, so accesses from the outside are given greater relative importance. This is desirable, since external opinion is thought as a more objective measure of relative importance.

4 Conducting the Experiment

We used the web server log from http://www.ceid.upatras.gr (Computer Engineering & Informatics Dept., University of Patras – the CEID site hereafter) in order to evaluate the results of the described algorithms. Experimental evaluation is based on the findings of the standard reactive web usage mechanisms that common web server software has; the log files. The most widespread log formats for HTTP servers are the W3C formats and which is more the extended [8] one. In this format among the recorded details, there are the IP address or DNS hostname of the user's host/ proxy and also the "referrer" URL, i.e. the page from which a request was initiated. Some commercial and to a lesser degree non-commercial sites also rely on cookies and server-supported proactive mechanisms for sessionization to identify user activity. In this context, proactive means that user identification is taken care of, while the user accesses the site. However, the largest proportion of sites still does not apply such techniques. As long as no proactive and privacy-compliant methods have become a standard among web servers, reactive heuristics will therefore remain essential for reliable usage analysis. We discuss the environment, the methodology and the results of the experimental procedure below. The reader may find extensive details of the experiment in our site[b], though this is not necessary in order to follow our discussion.

In order to facilitate our discussion, we will henceforth call, GKM the metric presented in [9] and TOP and PROB the metrics developed in Sections 3.2 and 3.3, respectively. We will use the same notation for the corresponding algorithms that incorporate the metrics.

4.1 Experimental Environment

We used a site that was hosted in a Sunfire 280r with a single Ultra SPARC processor at 900 MHz, using OS Solaris 8 and Web Server Apache 1.3.26. We obtained a web log covering 1.5 months (44 days), including 1,320,819 records (hits) and 3596 unique visitors (Feb 2002 – March 2002). We have also verified the first experiment using a second web log covering a period of 0.5 month (20 days) with 645,972 records (hits) and 3694 unique visitors (Sept. 2002). The results have been quite the same so we will refer to the first experiment in our discussion.

Our site structure eventually included 4 levels and 98 pages of interest. "Eventually" in this context means that our primary concern was to produce possible reorganization proposals for the structure of the main CEID site. As a consequence we had to make some judgment calls. First of all we excluded pages that they belonged to subordinate CEID sites (e.g. sites of CEID laboratories and certain faculty's and courses' private web sites, which we believe that they have their own structure). Moreover we dropped all responses of embedded resources such as images and non-data responses such as errors or search spiders & robots. We also considered different users as different clients based on IP addresses. This might interleave requests and sessions made by multiple users on the same

[b] The reader may optionally find full experimental resources, implementations and results through the following web site http://students.ceid.upatras.gr/~sakkopul/www2003.htm.

machine or users sharing the same IP address (e.g. proxy, RAS pool) but it does not change at all the fact that particular accesses have been made to particular pages. All this pre-processing produced a useful web log of 77,231 lines (hits) to the pages of interest (and 59,873 web log lines respectively for the second experiment).

4.2 Methodology & Data Preparation

The data we used to evaluate our algorithms represented a 6-9% of the initial web logs. At first we performed as discussed earlier an analysis of the structure of the site in order to recognize the pages of interest to us. In the sequel we implemented our proposed algorithms and metrics, the corresponding pre-processing procedures and the GKM algorithm and metric using the Mathworks Matlab v6.5 language of technical computing in order to perform the evaluation. We designed and implemented the pre-processing, parsing, distilling and extracting procedures from scratch in order to have full control of the filtering that was done. We discarded the responses with HTTP error codes [10] (i.e. status codes 4xx - 400, 401, 403, 404 etc). We also dropped all the redirection (i.e. status codes 3xx – 300, 301 etc), though they were not of substantial number as the structure and operation of the web server was quite stable. Furthermore we distilled the log in order to avoid potential web spiders' footprints (i.e. we excluded accesses that were done in the `robots.txt` dummy file). Finally we chose these entries of the log that were corresponding to pure HTML pages (i.e. extension htm & html) and we left out all additional files such as D-HTML part (i.e. extension css & js). We paid attention in the transformation of the URL to a standard format (e.g. common casing of host). We distinguished the incoming bookmark hits by matching the referrer with clients that accessed "bookmark this page" links (in the case of pages where users entered the site only). We had also to verify each log record for its compliance with the W3C extended log format (e.g. omit mistyped records due to web server restarting). Important, though, it may be that we used a frame-free version of the site in order to have more precise logs since it is shown that frame-based sites have serious impact on the mapping of the activities of each user to distinct actions [13].

4.3 Evaluation & Reorganization Lists

Before presenting the assessment of the results we would like to discuss the methods we followed in order to produce the reorganization proposals. Since the beginning we knew that the experimental site chosen was well structured and we expected that all of the algorithms would not provide a lot of reorganization proposals. The procedure that produces these suggestions is described hereafter.

Each of the algorithms, both our proposed ones and the GKM, produce a list of suggestions. We have designed the algorithms so as to propose for reorganization a list of web pages to the authors and administrators of the web site (who are the judging authority of all organizational aspects the web site). It is obvious that every page of interest in the site is assigned an *RA*. Our algorithms examine the current web site structure in order to discover possible abnormalities of the distribution of *RAs* in the site. We focused on tracking reasonable structural changes as far as the content of the pages is concerned. We performed thus, the examination in a local topological search by examining the *RAs* of the parents and children of each page in the site tree structure. Narrowing the examinations by focusing context-related site parts each time would behave in a similar manner (e.g. by applying the algorithms to context coherent sub-graph parts). In this case we incorporated in the algorithms the *RA* difference of the pages under examination. Moreover we wanted to depict the initial belief of a well-structured site. We inserted a context-based qualitative variable by considering that every parent is weaker than its child nodes in the case of an arithmetic operation tie (e.g. when the parent has exactly the same absolute accesses in the log with a child then it appears reasonable that the child is of more interest to the users) and therefore it has to be reorganized. We tested algorithms incorporating 5-10% weakness of the parent (0.9-0.95 of the parental RA) in different experiments and we did not receive any different proposals.

The following statement describes in plain terms the way the suggestions are produced.

$$if\,(Diff\,(RA_{child}\,,0.9RA_{parent}\,) > 0)\{proposal = true\}\,,$$

where *Diff* indicates the absolute difference in the *RA* between parent and child. We are using the quantity $Diff\,(RA_{child}\,,0.9RA_{parent})$ as a ranking function, presenting to the web master a proposed order of importance of the restructuring movements he should follow. The intuition behind this choice is that an inappropriate placement of a certain page must be apparent within the local context and the urge for reorganization can be reflected in the scale of local *RA* differences.

The proposal lists that are produced, are not limited in terms of length or popularity, because the at most judge, who is the potential web designer, needs to have a full perception of potential abnormalities in the site. Nevertheless we could further refine the listed results by the introduction of a notion of confidence. In that case confidence would describe the certainty that the algorithm has in the suggestions by limiting the output to those proposals that have a difference at least 30% of the largest one proposed already (related work [14] uses a support (similar to confidence) threshold of 5).

5 Assessment of results (metrics)

The GKM reorganization algorithm was a simple heuristic. They merely examined the relation between the Relative Access of a node and that of its parent; if a node has greater *RA* than its parent, then the two nodes are simply interchanged. Adopting this simple reorganization approach, we argue that the newly proposed metrics yield a better overall result.

5.1 Qualitative Assessment

The GKM metric provides a rough estimate of the multiplicative factor $a_{i,}$ The latter is affected to a large extent by the number of pages at the same level, regardless if they are siblings of page *i* or not. Thus, the *RA*s computed for sites of large width are considerably larger than the ones stemming from our metrics. Another point is that pages lying deep down the site hierarchy, get over more benefited, because of the tendency of web-sites to evolve like trees, i.e. the number of nodes at the same level grows as descending the site hierarchy. Since descendants of a node are over favoured, this can potentially lead to false reorganization estimates.

Both our approaches alleviate such phenomena, displaying a more balanced consideration of “node scoring”. Considering the TOP metric, we normalize the influence of both depth and width to the RA score. We further don’t take into account every single page of the same depth but only the siblings of a certain node, since the choice of a page among its siblings is what truly matters. The PROB metric behaves in much the same way. Our experimental results show that, figures computed by our metrics are, in about an order of magnitude, smaller than the respective ones of GKM, an indication that our techniques are more elaborate.

Another advantage of our approaches is the incorporation of bookmarks. Bookmarks are an important indicator of page quality. Thereby, we estimate that the number of bookmarks on a page, can provide some sort of quality estimate. This was our motivation for incorporating the number of bookmarks. We further extend the simple consideration of bookmarks by assigning a weight, according to the depth of the bookmarked page.

Assessing our proposals, PROB is expected to give more accurate results in general sites, where a large number of crossing links exist, leading to a general view of a site as a DAG. Albeit, this was not the case in our site, which exhibited an almost "perfect" tree organization, therefore, both our metrics give more or less the same results.

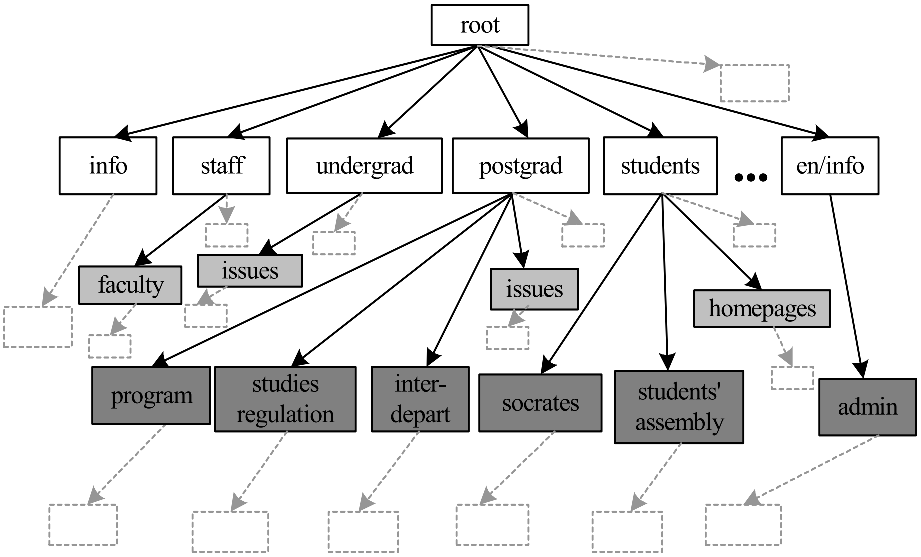

Figure 2. A fragment of CEID site

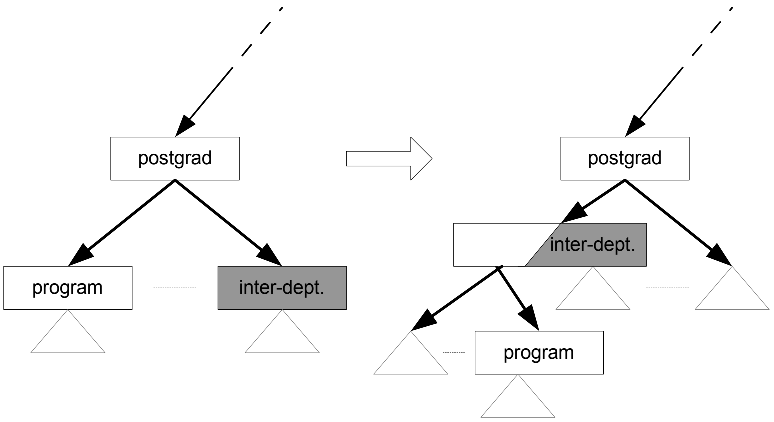

Figure 3. Local structure changes. The instance to the left represents the situation afterwards. The absorbed node is shaded. Subtree nodes are drawn in gray borders.

5.2 Analysis of Experiments

Before proceeding with the discussion, it is important to mention that after we had gathered our log files, the CEID site went under a scheduled redesign by the webmasters, a process *independent* of our work. It is worthwhile, however, to observe how this independent piece of work is related to our results. Overall, one may recognize the reconstructured nodes of the new CEID site in the light grey filling colour of Figure 2 (the proposals by our algorithms may be found in *Table 1* & *Table 2*). These nodes substituted simple navigation index nodes that were present in the same place of the corresponding website sections (i.e. faculty, postgraduate & undergraduate studies and student users).

	Proposals[c]	**Differentiations**		
		PROB–TOP	GKM–TOP	GKM–PROB
1	(http://www.ceid.upatras.gr/postgraduate/program.htm, http://www.ceid.upatras.gr/postgraduate/index.htm)	√	√	-
2	(http://www.ceid.upatras.gr/postgraduate/inter-dep.htm, http://www.ceid.upatras.gr/postgraduate/index.htm)	√	√	-
3	(http://www.ceid.upatras.gr/students/socrates.htm, http://www.ceid.upatras.gr/students/index.htm)	×	√	√
4	(http://www.ceid.upatras.gr/students/assembly.htm, http://www.ceid.upatras.gr/students/index.htm	×	√	√
5	(http://www.ceid.upatras.gr/postgraduate/regulation.htm, http://www.ceid.upatras.gr/postgraduate/index.htm)	×	√	√
6	(http://www.ceid.upatras.gr/en/info/administration.htm, http://www.ceid.upatras.gr/en/info/index.htm)	×	√	√

Table 1. Differentiated proposals: ‘x’, indicates that these pairs were not suggested by either algorithm, ‘√’ indicates a difference and ‘-‘ no difference.

Our aim will be, at first place, to interpret the differences between the three algorithms, as those stemmed from the experiments. In the first round, we extracted from the processing of the log file a set of suggestions. *Table 1* summarizes the differences in the proposals of the three observed metrics, TOP, PROB and GKM. Regarding the table contents, we adopt the notation X – Y, to denote the proposals of metric X that are not included in the suggestions of metric Y. The differences listed in the *Proposals* column are of the form (*page*, *page's parent*). Recall that we are interested in pages that have *greater* Relative Access than their parent, a phenomenon that these pairs exhibit. This anomaly yields a potential reorganization proposal as the ones listed in *Table 1*.

The first two proposals (nodes *program* and *inter-departmental* and their common parent *postgrad* in Figure 2) are common to PROB and GKM and do not appear in TOP. The independent redesign of the CEID site treated those suggestions under a localized rearrangement plan, supporting this proposal independently. In particular an intermediate node was inserted between the *program* page which, in addition, absorbed the *inter-departmental* page (see Figure 3).

Another observation concerns the fourth proposal. The child page involved here was totally removed from the new site. Nonetheless, the same proposal was treated as significant by the GKM metric. Note though, that neither of our proposed metrics includes this proposal. Furthermore, the rest of the proposals on *Table 1* – namely 3, 5, 6- fall into the so called *irrational* category, rearrangements that are impossible to be performed. An elucidating example would be the interchange of, say, the

[c] The actual links have been translated in order to increase comprehension. One should use the given substitutes in order to have the actual site links. Substitutions: **metaptyxiaka** for *postgraduate* (issues), **pms** for (postgraduate) *program*, **diatmhmatika** for *inter-dep*(artmental), **kanonismos** for (studies) *regulation*, **syneleysh** for (students') *assembly*

webpage of undergraduate issues with its parent, the academic issues directory; what differentiates them from the previous ones is that they are only proposed by GKM. The above discussion substantiates the qualitative arguments of Section 5.1. To put it in other words, the methods we introduce have been proved to be more elaborate and subsequently produce less false estimates. This property is important in the case of well-structured sites, such as the site we were experimented on, where the webmaster must examine carefully every potential proposal.

	Proposals	**Substitutions**[d]
1	(http://www.ceid.upatras.gr/undergraduate/issues.htm, http://www.ceid.upatras.gr/undergraduate/index.htm)	**proptyxiaka** is substituted by undergraduate and **themata** by (undergraduate) issues
2	(http://www.ceid.upatras.gr/staff/faculty.htm, http://www.ceid.upatras.gr/staff/index.htm)	**prosopiko** is substituted by staff
3	(http://www.ceid.upatras.gr/students/students.htm, http://www.ceid.upatras.gr/students/index.htm)	
4	(http://www.ceid.upatras.gr/postgraduate/themata.htm, http://www.ceid.upatras.gr/postgraduate/index.htm)	**metaptyxiaka** is substituted by postgraduate and **themata** by (postgraduate) issues

Table 2. Proposals for the modified site.

We also needed verification that our method was sound. Therefore we mirrored the site in a local computer and modified its structure in a way that some highly popular pages that were directly linked from the homepage but also had an alternative parent, were cut off from the homepage. This setting is depicted in Figure 2, in the case of the light grey nodes. We ran our experiments using the same log file. Our expectation is that since these pages were highly popular this should be apparent in the proposals. In the spirit of Section 4.3, proposal lists were extracted, the top results of which are displayed in *Table 2*.

The proposals format is identical to that of *Table 1*. No extra information is provided due to the fact that the proposals list and the order of appearance in it is identical to all the three metrics. A noticeable fact here is that all the pages, which were deliberately modified, appear as top suggestions for reorganization, exactly as expected to be

At this point we can infer that the metrics, we described, are capable of foreseeing, to a significant extent, the changes that seem appropriate to a web site, according to its visit patterns. One should also keep in mind that the site at our disposal was well structured, not likely thus to generate several proposals

5.3 Information Extracted from Charts

We believe that meaningful information can be drawn out of a chart displaying the *RA*s, of every page, such as the one depicted in Figure 4.

Imagine a perfectly organized site, i.e. a site with no need for reorganization. Suppose that we number the nodes of the site tree in a Breadth First Search (BFS) order. It is commonsense to believe that in this perfect setting the following conditions will hold:

- All the nodes within the same level, retain similar *RA* values, with the presence of some fluctuations

[d] This column indicates URL substitutions as those needed in Table 1

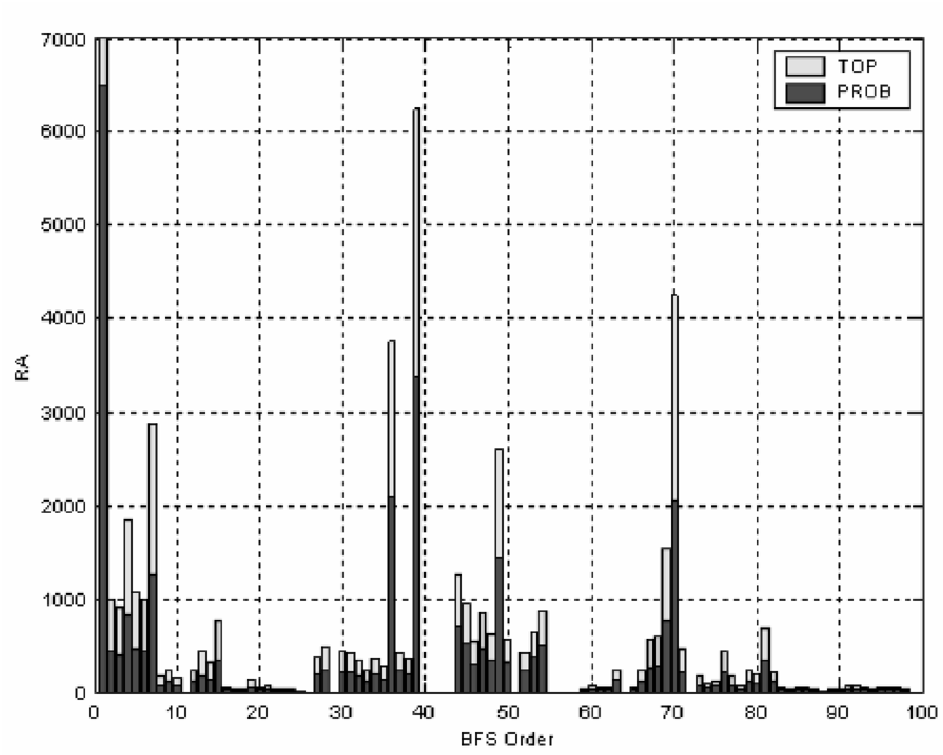

Figure 4. The RA accesses of the CEID site pages in BFS order

- The *RA* values decrease rapidly as descending from a level to the immediate next one. That is the *RA* score continuously increases as moving towards the root.

The above conditions are best illustrated in Figure 5. Recall that we have postulated a BFS order, the diagram, thus, is expected to have a staircase-like form.

Contrasting Figure 4 with Figure 5, it can be observed that *RA*s in our site don't act in this way an indication for reorganization and this case is presumably the general one. A rule of thumb can arise thus; define a threshold for each tree level and let the RA peaks trespassing this threshold correspond to reorganization candidates.

5.4 ReOrganization Proposals

All discussions about site reorganization describe procedures to produce a list of web pages that have to somehow be rearranged within a site. In contrast to dynamic proposals (e.g. e-commerce recommendation lists, pop-up agents etc) we are talking about wide and perhaps long-term, structural and contextual changes. In most cases the implementation of these changes does not include any notion of an engineering approach or it is rather simplistic. In the web site reorganization "literature" (see citations in Section 1) a main stream of the presented suggestions includes the transformation of

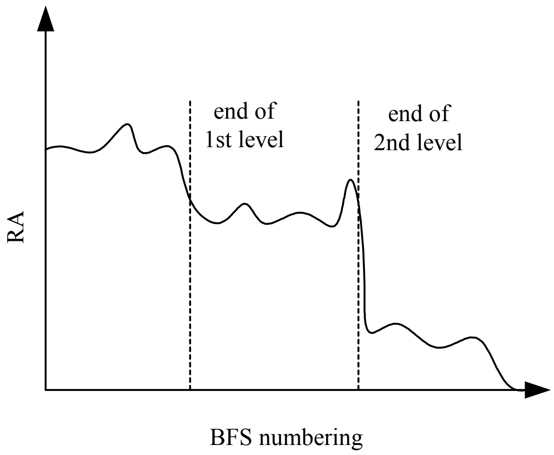

Figure 5. The staircase formed by the RAs in a fictional, perfectly organized website

the proposal list to a simple recommendation list of “hot” or “most wanted” link (usually in an HTML frameset). Another popular approach is the over-simplified swap between pages, case which is particularly used, having in mind the theoretical swap of tree nodes and it does not include implementation details. Other similar cases include the insertion of a new hyperlink between the pages proposed for reorganization either followed by deletion of existing site structure links or by ignoring them.

Alternatively, substitution (i.e. merging) of two or perhaps more pages is proposed; that option is especially applicable to pages concerning items of information with close similarity and small differences in their *RAs*. Furthermore, another relevant case is the use of preview methods (i.e. inline frames-iframes [11]) so as to present only a part of the page under reorganization in the proposed site node.

In our experiments we have observed that there is a need to associate the proposals with specific kinds of reorganization methods, since otherwise they may harm contextual coherence. We propose the distinction between proposals of strong and weaker *confidence* by splitting the suggestions into the two following groups (see *Table 3*). However, a reorganization proposal framework is needed to apply such changes accurately and automatically (see an early approach in [6]).

Proposal Confidence	**Web Technique**
Strong	Merging, preview, hot link lists
Weak	Extra link (with or without deletion of previous link connections)

Table 3. Reorganization techniques and proposals

6 Site Design, User Behaviour and Proposed Redesigns

Undoubtedly, it would be interesting to discover the relation, if any, between correct design practices and reorganization proposals, or more precisely, the lack of them. In other words, it is questionable whether it is likely that in a well designed website, hardly any need for reorganization occurs.

In such a site, the actual design decisions are mostly dictated by two factors:

1. Designer's intuition,
2. Human Computer Interaction, design guidelines.

In any site, even in a "well-designed" one, it is not easy to conclude, whether it will be susceptible to restructuring. The latter depends completely on users' behaviour.

Our experience from a well-designed site, such as the one we experimented on, suggests that only a few proposals usually result, some of which may be characterized as *irrational* and therefore should be judged by the web author (cf. Section 5.2). As an illustrating example, it is most unlikely that the webmaster will decide to exchange places between the page that contains information about the postgraduate programme and the index page for postgraduate issues, as suggested by proposal no. 1 in *Table 1*. Albeit, he may decide to perform a merger between these two pages, even a partial one.

On the other hand, when the designer's ultimate choice has led to false design decisions, this can be reflected in the access patterns of the site and consecutively in the *RA*s. It would be interesting to observe and assess the proposed redesigns in such a site, however we didn't have such a site in our disposal.

7 Conclusions and Future Work

This work aims to provide refined metrics, useful techniques and the fundamental basis for high fidelity website reorganization methods and applications. We have introduced metrics & methods that may considerably improve website structure to the users' navigational favour. In particular, we proposed and tested two new metrics that proved efficient in praxis. A comparative analysis involving a previously known metric is also performed. We believe that our approach may easily be incorporated into the series of web traffic analysis tools, becoming in this way the cornerstone of website optimization.

Future steps include the description of a framework that it would evaluate the combination of reorganization metrics with different sets of redesign proposals. We also consider as open issue the definition of an overall website grading method that would quantify the quality and visits of a given site before and after reorganization, justifying thus the instantiation of certain redesign approaches.

Acknowledgements

The authors would like to thank George Oikonomou, for his invaluable help during the project. We would also like to thank an anonymous referee for his useful comments on the improvement of our presentation.

References

1. Berkhin, P., Becher, J.D. & Randall, D.J.: Interactive path analysis of web site traffic. *Proceedings of KDD01* pp. 414-419, 2001.

2. Bose, P., Kranakis, E., Krizanc, D., Vargas Martin, M., Czyzowicz, J., Pelc, A. & Gasieniec, L.: Strategies for Hotlink Assignments. In *Proceedings of the International Symposium on Algorithms and computation*, (ISAAC 2000), pp 23-34, LNCS 2223, Springer Verlag, 2000

3. Botafogo, R.A., Rivlin, E. & Shneiderman, B.: Structural Analysis of Hypertext: Identifying Hierarchies and Useful Metrics, *ACM Transactions on Information Systems,* vol. 10, no 2, pp. 142-180, April 1992

4. Brin S. & Page L., The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30(1-7): 107:117, 1998

5. Chen Ming-Syan, Jong Soo Park, & Philip S. Yu. Data mining for path traversal patterns in a web environment. In *Proc. of the 16th International Conference on Distributed Computing Systems*, pp. 385-392, 1996.

6. Christopoulou, E., Garofalakis, J, Makris, C., Panagis, Y., Sakkopoulos, E. & Tsakalidis, A. Automating restructuring of web applications, poster presentation in *ACM HT 02*, (available through the following link http://mmlab.ceid.upatras.gr/HT02/ht2002.pdf).

7. Drott M.C. Using web server logs to improve site design *Proceedings of ACM SIGDOC 98* pp.43-50, 1998.

8. Extended Log File Format W3C. http://www.w3.org/pub/WWW/TR/WD-logfile.html.

9. Garofalakis, J.D., Kappos, P. & Mourloukos, D.: Web Site Optimization Using Page Popularity. *IEEE Internet Computing* 3(4): 22-29 (1999)

10. Hypertext Transfer Protocol, RFC 2616, W3C. http://www.w3.org/Protocols/rfc2616/rfc2616.html.

11. Inline Frames, HTML 4 W3C recommendation http://w3c.org/TR/REC-html40/present/frames.html

12. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. *JACM 46(5)*: 604-632 (1999)

13. Spiliopoulou, M., Mobasher, B., Berendt, B. & Nakagawa, M.: A framework for the evaluation of session reconstruction heuristics in the web usage analysis. *INFORMS Journal on Computing*. *Special Issue on Mining Web-Based Data for E-Business Applications*, Vol. 15 No. 2, 2003.

14. Srikant, R., Yang, Y.: Mining web logs to improve website organization. In *Proceedings of the WWW10*, Hong-Kong, pp 430-437, 5/2001

15. Zhou Baoyao, Jinlin Chen, Jin Shi, HongJiang Zhang, Qiufeng Wu: Website link structure evaluation and improvement based on user visiting patterns**.** In *Proceedings of the 12th ACM Conference on Hypertext* (HT01), pp. 241-244, 2001

16. Web reference:Analog. http://www.analog.cx

17. Web reference:SurfStats http://www.surfstats.com

18. Web reference:Web Trends http://www.webtrends.com

19. Web reference:WebLogs http://www.cape.com