Coupling Logfile Analysis and Content Management Systems for Improved Information Architecture

Roland Mücke
IMISE, University of Leipzig
Härtelstraße 16-18
04107 Leipzig, GERMANY
+49 341 97 16165
roland.muecke@imise.uni-leipzig.de

Matthias Löbe
IMISE, University of Leipzig
Härtelstraße 16-18
04107 Leipzig, GERMANY
+49 341 97 16113
matthias.loebe@imise.uni-leipzig.de

ABSTRACT
In this paper, we describe a methodology for coupling the analysis of web server logfiles with content management systems. This coupling offers benefits for the analysis of user behavior, the detection of issues within a website, and the feedback of analysis results to the website author. Therefore, the quality of websites can be enhanced by this methodology.

Categories and Subject Descriptors
H.5.4 [**Information interfaces and presentation**]: Hypertext/Hypermedia – *Navigation, User issues*

General Terms
Algorithms, Management, Measurement

Keywords
Logfile Analysis, Content Management Systems, User Behavior

1. INTRODUCTION
Analyzing web server log files and interpreting access statistics is one aspect of running a website. There are numerous reasons to do this. For instance, one may want to measure how successful a website is, i.e. analyze how many users get attracted to it, how many pages they visit or how long they stay [7]. These questions concern how good the website actually is or was, measuring it with numbers of page impressions as a rather simple metrics. But for improving a website, this kind of data will not help that much [4]. Instead, information about the user's interests and desires are needed. Knowledge about problems of a website, especially with respect to finding desired content and navigation on that site, is very valuable, too. Giving a user what he wants and letting him quickly and comfortably find what he is looking for will have impact on his positive perception of the website and finally make it more successful [3].

To address aspects of user behavior, additional information about the accessed pages is required for the analysis of logfiles, like keywords describing the content or the intended audience of a page. Such metadata can be provided by a content management system that is used to manage the website and its assets. The enrichment of logfiles with metadata is one part of the coupling between logfile analysis and content management systems.

The second part of the coupling addresses feedback to the content management system. The results of logfile analysis may reveal opportunities for improvements, like the promotion of important pages that are accessed infrequently or the detection of problems, like pages which cause users to get lost [2]. Based on such findings, changes to single web pages or a website's navigational structure may be suggested. These suggestions can be fed back into the content management system where they can be directly presented to editors and site managers.

In this paper we outline a methodology for coupling logfile analysis and content management systems. In section 2, the aspects of improving the information architecture of a website that are relevant to our methodology are described. After that, we introduce the metadata required for the enhanced logfile analysis in section 3. Then the methodology is presented in section 4, followed by a discussion and an outlook on future work in section 5. Finally, we conclude this paper in section 6.

2. IMPROVING INFORMATION ARCHITECTURE
Improving a website is a continuous process. Firstly, this includes the adaption and refinement of contents in which users are interested. Common improvements are the addition of pages that cover special topics or the enhancement of particular functionalities. Another type of improvement is the promotion of certain topics that are important for the users – at least from the perspective of the website owner. An example of this kind is a web page with important information about some service the website provides. If only few users access this page but everybody should, because it greatly enhances that services use, then the link to this page should be presented in a way that attracts more people. This can be achieved by placing it on a more prominent position or by providing it with some "graphical sugar" like an icon or a special font style.

Another aspect of improving a website is to eliminate problems with its navigation. The "Lost in hyperspace" syndrome [5] is an issue which plagues people since the early days of hypertext systems. To deal with this issue, one may rework certain pages that are identified as causing users to get lost easily or one may reorganize the navigational structure of the website in general such that it is more compliant with the way users usually move around on it.

Altogether, knowledge about users is highly valuable in order to achieve improvements. When we talk about "users", we do not mean one stereotype but individuals with their own goals and motivations to access a website. Therefore, one should not lump together all users and build a "one size fits all" website. Taking into account the existing diversity of the users will help to shape a website such that it offers its contents in a way more specific to the users needs [1]. On a web portal where every user has a login and is personally known to the website, it is relatively easy

to customize content and navigation. A survey of user interests or background knowledge concerning a website's topics can be carried out, e.g. during account setup. This explicit information can then be used to guide the user to that contents he is or may be interested in. Of course, not all websites can be portals. If one offers contents and services to an anonymous community of users it is hard to identify certain patterns of interests and background knowledge. They have to be inferred from the traces the users leave during their visits.

3. REQUIRED METADATA

Techniques like analyzing the logfiles of the web server software or tracking users with "web bugs" offer broad information about user behavior like which pages are used as entry or exit pages or which pages are accessed most. However one cannot infer much about the user's interests or background knowledge. It is also hard to identify pages that cause users to get lost on the website. The reason for this is that the information provided by web server logfiles and tracking systems is not sufficient for these questions. In these files only data about the requested URLs are covered but nothing is said about the contents and purpose of the pages identified by these URLs. Therefore the logfiles need to be augmented with additional data about the accessed web pages. Those metadata should cover the aspects of user interests, background knowledge and problems with website navigation mentioned before. The following metadata are proposed to enrich logfiles:

- **Audience** – a classification of users for which a page is intended. It reflects the assumed background knowledge of the readers and their interests. For example, a medical website may have the target audiences "patients" and "physicians". Pages for patients offer a simpler language, whereas pages for physicians contain more in-depth information.

- **Content classification with keywords** – a simple description of the contents provided by each page that helps to relate several pages with the same topic to each other. On a medical website this could be something like "therapy", "diagnosis" or "drug".

- **Language** – if there are different language versions of a page, this value indicates the language a user seems to prefer.

- **Navigational type** – the function of the page for navigating around the website [6]. Those could be "homepage", "content page" for pages with the actual content, "overview page" for pages with a short description of a topic and a list of links to content pages, and "utility page" for sitemaps and search pages.

- **Importance** – a value for rating a page's importance from the perspective of the website owner. In the example of a medical website, some pages about a new therapy may be rated very important. This metadata may be combined with the *target audience* and *content classification* to get a more specific description of the content of a page.

- **Links from other pages** – a list of all pages of the website that link to a page. The more pages that link to a certain page, the better it is accessible.

There are several places where these metadata may be stored and retrieved during logfile analysis. They can be put directly into the HTML code of web pages. This is rather simple but publishes information like the importance rating of a page which should not be shown to the users. An external metadata store like an RDF database could be more adequate, but setting up one just for logfile analysis might be too costly because of the additional software system that needs to be maintained. For larger websites with lots of pages and many people working on them, content management systems (CMS) are often used. For its management duty, the CMS has to store information about the pages, so it would be a convenient approach to use it as the source for the metadata for logfile enrichment. That is why it is suitable for coupling it with logfile analysis.

4. METHODOLOGY

The methodology for coupling the analysis of web server's logfiles and a content management system consists of four steps. As shown in figure 1, they form a process which starts with accessing metadata and logfiles and ends with the manipulation of content and navigation. In the first step, the logfiles from the web server are enriched with metadata generated by means of

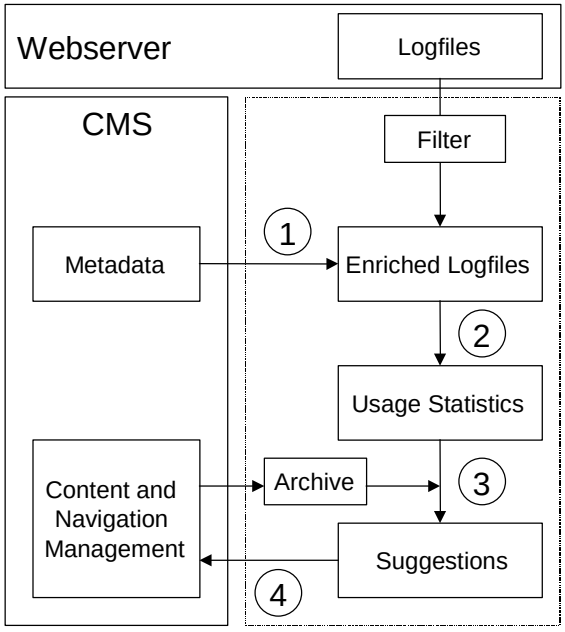

Figure 1: Process with four steps: (1) Enrichment, (2) Analysis, (3) Generation, (4) Presentation and Application

the CMS. In step 2, they are analyzed and usage statistics are created. Those contain information about user interests and issues with the navigational structure of the website. Based on this information, suggestions for improvements are generated in the third phase. Finally, these suggestions are presented to website editors and possibly applied to the content or navigational structure.

This process may be automated to a certain degree so that it can be executed in regular intervals. This can have great impact on the costs of continuously reviewing and improving the website.

Instead of a major annual web usage analysis which requires a lot of time and the knowledge of experts in information architecture to find simple and difficult problems with the website, the automated process already handles the simple flaws, so that the expensive experts can focus on the difficult ones. This reduces costs and at the same time improves the website.

All four steps are explained in detail in the next subsections.

4.1 Enriching Logfiles with Metadata

In the first step, the logfiles from the web server are prepared and the metadata are retrieved from the CMS. Those requests not relevant for the analysis are filtered out, e.g. unsuccessful requests and utility files like graphics. Furthermore the user sessions are identified as usual using IP addresses, referrers, user agent information and time frames. Afterwards, a list of unique URLs is created. These URLs represent the web pages that were accessed and that shall be analyzed. They are used to identify the pages in the content management system. This assumes that the CMS is able to map URLs to its internal representation of web pages which is not always the case. If rewriting of URLs is performed on the web server, beyond the control of the CMS, this mapping may be impossible, so these rewrites have to be set off.

Now metadata can be retrieved from the pages in the CMS. For this purpose a public API to the CMS is very valuable. If such an API does not exist one has to access the database structure of the CMS directly, which might be problematic. Fortunately, most CMS have the goal of being highly customizable and extendable by their users so that a public API is often available. With the specification of the Java Content Repository (JCR), at least in the Java world there exists a standardization effort so that generic access to the metadata of web pages can be provided.

For the analysis of the logfiles in the next step of the methodology, metadata described in section 3 are required. We assume that they are acquired during the creation of the web pages or they can be automatically determined by the CMS software. Since the following steps heavily depend on these metadata, it is very important that they are complete and accurate.

As a result of this step, the preprocessed logfiles and the metadata of every web page are available for analysis.

4.2 Logfile Analysis

Now the analysis of the user behavior can be performed. In this paper, we focus on three aspects, but further analysis strategies may be developed in the future as described in section 5. These three aspects concern the topics of interest for the audiences, pages that cause problems with orientation on the website, and pages that are rarely visited contrary to their designated importance.

4.2.1 Identifying user interests

From the distribution of a keyword among the pages accessed during one visit, one may infer how much that user was interested in the topic associated with that keyword. Requesting just a single page about a topic has almost no reliable indication that one is actually interested in it. But the higher the number of pages about a topic is, the greater is the probability of an interest in that topic. For example, if eight out of ten pages of a visit are associated with the keyword “therapy”, than one can quiet reliably assume that this user was really looking for information about a therapy.

The evidence can be increased if the time the user spent on each page is included in the calculation. A page where he stayed for one minute has far more impact than a page that was accessed just for five seconds. Chances are high that the first page was really read whereas the second one was just scanned and it was quickly decided that it is of no interest.

In the same way it may be identified whether a user belongs to one of the audiences the website owner focuses on. If a great percentage of the pages the user visited are intended for physicians, then he may be a physician himself or may have at least enough medical knowledge to understand the content. This assumption can be made more reliable if the “degree of difficulty” of a page’s content is stated. A page full of medical terms may indeed only be understandable to a physician so it is unlikely that a patient will spend a long time on it.

4.2.2 Identifying confusing pages

During a visit, a user accesses different types of pages. He may start with the homepage, navigate through “overview pages” or use a “utility page” like the sitemap or the search function of the website in order to finally access certain “content pages”. If this kind of conversion towards content does not occur, this might be interpreted as that the user has got lost. For example if the sitemap or a search page is accessed after the user wandered through “overview pages” and “content pages” for a while, it shows that he might has changed his search strategy because he could not find what he was looking for. Returning to the homepage of the website may have the implication that the user starts over again, which also means that he was not able to locate the desired content.

We rate this degree of lostness by counting page transitions towards or within “content pages” versus those transitions away from the content, i.e. to the homepage or to utility pages. If that statistic indicates that the user seems to be lost, one can detect the pages responsibility for the disorientation by looking at the first pages accessed during the visit. These pages have a high impact on the overall orientation of the user and may therefore have confused him so that he got lost.

4.2.3 Rating the success of important pages

If the importance of pages is specified in the metadata, it can be automatically rated whether such a page gets the desired attention compared to similar pages with the same topic or same audience. If this is not the case, possible reasons for this can be found by further analyzing the metadata of that page. Two of those reasons that can be checked are the number of pages that link to the problematic page and the position of those links, especially in navigation menus. Placing links on a more prominent position may be one of the changes suggested to the website editors.

4.3 Generating Suggestions

Based on the results of the logfile analysis some suggestions for improvements or changes of the website or single pages can be produced. Of course such suggestions will not compete the suggestions an expert in information architecture or website usability would make. But that is not our main goal anyway. Instead, the automatic generated suggestions should address the

simple and obvious flaws, and this can be done in regular intervals without high costs caused by involving an expert.

Generated suggestions might have the form of a list of topics that should be covered in more detail or a list of pages that should be reworked because they seem to confuse the users instead of guiding them. For that generation, a library of suggestions can be used, which may be extended by further strategies for improvements that might be useful for a particular website.

If the process of logfile analysis and the generation of suggestions is performed regularly, it can occur that suggestions are created over and over again, and always rejected by website editors. To avoid this annoyance, the acceptance or rejection of a suggestion would be logged in order to supress the generation of new suggestions in the next run of the process. In this way, for example, the reworking of an overview page is not proposed again although a positive effect after a rework remains to be seen.

4.4 Presentation and Application

Finally, the suggestions are presented directly in the content management system so that editors can easily access those suggestions and let them influence their work on web pages or the navigational structure of the website. The integration of the results of the logfile analysis into that system where the content and the entire website is managed may improve the productivity of website management because editors just need to use one tool, the CMS, and the time-consuming task of transferring suggestions from a separate tool for logfile analysis to the CMS can be omitted.

The acceptance or rejection of a suggestion is logged in an archive (see Figure 1) and influences the generation of such a suggestion in future iterations of the analysis process.

For this second part of the coupling (the first one was the enrichment of logfiles with metadata) again some kind of interface to the CMS is required. But this time an interaction with the user interface of the CMS is needed. So it may be difficult to obtain such an API because it accesses functionalities specific to the CMS which might not be intended to be public accessible, especially on commercial products that are not open source. Hence, systems which focus greatly on customization and flexibility might have an advantage when it comes to this part of the methodology.

5. DISCUSSION AND OUTLOOK

Although the methodology outlined in this paper offers several benefits, there are some difficulties in deploying it in real world scenarios. Besides the problem of the availability of an API for the CMS powerful enough to fully implement the coupling, there are practical issues concerning the acquisition of metadata and their modification over time. The importance of metadata is often not recognized by website editors and thus they are not acquired accurately. Furthermore, metadata changes are not always properly logged so that analyzing older logfiles leads to incorrect results.

The techniques for analyzing user behavior and generating suggestions for improvements described in sections 4.2 and 4.3 show only some of the opportunities what can be done if additional metadata are available for analysis. More techniques may be developed that use other metadata or combine them in other ways. For example, the use of non-personal logins for secured areas of a website may heavily influence a user's assumed membership of an audience. Also the partitioning of user sessions into phases with different interests or goals may be subject of future research.

6. CONCLUSION

In this paper we have presented a methodology for coupling logfile analysis and content management systems. The methodology consists of four phases. First, the enrichment of web server logfiles with metadata from the CMS was described. After that the enriched logfiles are analyzed and, based on the results, suggestions for improvements are generated. Finally these suggestions are presented in the CMS and applied by website editors.

This coupling has several benefits. With the use of metadata about the web pages it is possible to get a deeper understanding of user behavior. Also problems as well as opportunities for improvements can be detected. Further on, the automation of the process makes it possible to execute it in regular intervals. Thus a continuous review of the website can be performed and simple flaws be detected without the involvement of an expensive expert which then can concentrate on the bigger problems of website evolution.

7. REFERENCES

[1] Benyon, David; Höök, Kristina: Navigation in Information Spaces: supporting the individual. INTERACT '97: Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (1997), 39-46

[2] Dhyani, Devanshu; NG, Wee K.; Bhowmick, Sourav S.: A survey of Web metrics. ACM Computing Surveys 34 (2002), December, no. 4, 469–503.

[3] Donahue, George M.: Usability and the Bottom Line. IEEE Software 18 (2001), January/February, no. 1, 31–37

[4] Kohavi, Ron: Mining E-Commerce Data: The Good, the Bad, and the Ugly. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2001), 8-13

[5] Otter, Malcom; Johnson, Hilary: Lost in hyperspace: metrics and mental models. Interaction with Computers 13 (2000), September, no. 1, 1–40

[6] Pirolli, Peter; Pitkow, James; Rao, Ramana: Silk from a sow's ear: extracting usable structures from the Web. Proceedings of the SIGCHI conference on Human factors in computing systems (1996), 118–125

[7] Woon, Y.-K. Wee-Keong Ng Ee-Peng Lim: Evaluating Web Access Log Mining Algorithms: A Cognitive Approach. WISE Workshops (2002), 217-222