A Website Mining Model Centered on User Queries

Ricardo Baeza-Yates[1],[3],[2] and Barbara Poblete[2],[3]

[1] ICREA, Barcelona, Catalunya, Spain
[2] Center for Web Research, CS Dept., University of Chile
[3] Web Research Group, Dept. of Technology, Univ. Pompeu Fabra, Barcelona, Spain
{ricardo.baeza,barbara.poblete}@upf.edu

Abstract. We present a model for mining user queries found within the access logs of a website and for relating this information to the website's overall usage, structure and content. The aim of this model is to discover, in a simple way, valuable information to improve the quality of the website, allowing the website to become more intuitive and adequate for the needs of its users. This model presents a methodology of analysis and classification of the different types of queries registered in the usage logs of a website, such as queries submitted by users to the site's internal search engine and queries on global search engines that lead to documents in the website. These queries provide useful information about topics that interest users visiting the website and the navigation patterns associated to these queries indicate whether or not the documents in the site satisfied the user's needs at that moment.

1 Introduction

The Web has been characterized by its rapid growth, massive usage and its ability to facilitate business transactions. This has created an increasing interest for improving and optimizing websites to fit better the needs of its visitors. It is more important than ever for a website to be found easily in the Web and for visitors to reach effortlessly the contents they are looking for. Failing to meet these goals can result in the loss of many potential clients.

Web servers register important data about the usage of a website. This information generally includes visitors navigational behavior, the queries made to the website's internal search engine (if one is available) and also the queries on external search engines that resulted in requests of documents from the website, queries that account for a large portion of the visits of most sites on the Web. All of this information is provided by visitors implicitly and can hold the key to significantly optimize and enhance a website, thus improving the "quality" of that site, understood as *"the conformance of the website's structure to the intuition of each group of visitors accessing the site"* [1].

Most of the queries related to a website represent actual information needs of the users that visit the site. However, user queries in Web mining have been

studied mainly with the purpose of enhancing website search, and not with the intention of discovering new data to increase the quality of the website's contents and structure. For this reason in this paper we present a novel model that mines queries found in the usage logs of a website, classifying them into different categories based in navigational information. These categories differ according to their importance for discovering new and interesting information about ways to improve the site. Our model also generates a visualization of the site's content distribution in relation to the link organization between documents, as well as the URLs selected due to queries. This model was mostly designed for websites that register traffic from internal and/or external search engines, even if this is not the main mechanism of navigation in the site. The output of the model consist of several reports from which improvements can be made to the website.

The main contributions of our model for improving a website are: *to mine user queries within a website's usage logs*, *obtain new interesting contents to broaden the current coverage of certain topics in the site*, *suggest changes or additions to words in the hyperlink descriptions*, and in a smaller scale *suggest to add new links between related documents and revise links between unrelated documents in a site*.

We have implemented this model and applied it on different types of websites, ranging from small to large, and in all cases the model helps to point out ways to improve the site, even if this site does not have an internal search engine. We have found our model specially useful on large sites, in which the contents have become hard to manage for the site's administrator.

This paper is organized as follows. Section 2 presents related work and section 3 our model. Section 4 gives an overview of our evaluation and results. The last section presents our conclusions and future work.

2 Related Work

Web mining [2] is the process of discovering patterns and relations in Web data. Web mining generally has been divided into three main areas: *content mining*, *structure mining* and *usage mining*. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in a website:

Content: The real data that the website was designed to give to its users. In general this data consists mainly of text and images.

Structure: This data describes the organization of the content within the website. This includes the organization inside a Web page, internal and external links and the site hierarchy.

Usage: This data describes the use of the website, reflected in the Web server's access logs, as well as in logs for specific applications.

Web usage mining has generated a great amount of commercial interest [3, 4]. The analysis of Web server logs has proven to be valuable in discovering many

issues, such as: if a document has never been visited it may have no reason to exist, or on the contrary, if a very popular document cannot be found from the top levels of a website, this might suggest a need for reorganization of its link structure.

There is an extensive list of previous work using Web mining for improving websites, most of which focuses on supporting adaptive websites [5] and automatic personalization based on Web Mining [6]. Amongst other things, using analysis of frequent navigational patterns and association rules, based on the pages visited by users, to find interesting rules and patterns in a website [1, 7–10]. Other research targets mainly modeling of user sessions, profiles and cluster analysis [11–15].

Queries submitted to search engines are a valuable tool for improving websites and search engines. Most of the work in this area has been directed at using queries to enhance website search [16] and to make more effective global Web search engines [17–20]. Queries can also be studied to improve the quality of a website. Previous work on this subject include [21] which proposed a method for analyzing similar queries on Web search engines, the idea is to find new queries that are similar to ones that directed traffic to a website and later use this information to improve the website. Another kind of analysis, is presented in [22] and consists of studying queries submitted to a site's internal search engine, and indicates that valuable information can be discovered by analyzing the behavior of users in the website after submitting a query. This is the starting point of our work.

3 Model Description

In this section we will present the description of our model for mining website usage, content and structure, centered on queries. This model performs different mining tasks, using as input the website's access logs, its structure and the content of its pages. These tasks also includes data cleaning, session identification, merging logs from several applications and removal of robots amongst other things which we will not discuss in depth at this moment, for more details please refer to [23–25]. The following concepts are important to define before presenting our model:

Session: A session is a sequence of document accesses registered for one user in the website's usage logs within a maximum time interval between each request. This interval is set by default to 30 minutes, but can be changed to any other value considered appropriate for a website [23]. Each user is identified uniquely by the `IP` and `User-Agent`.

Queries: A query consists of a set of one or more keywords that are submitted to a search engine and represents an information need of the user generating that query.

Information Scent: IS [26] indicates how well a word, or a set of words, describe a certain concept in relation to other words with the same semantics.

For example, polysemic words (words with more than one meaning) have less IS due to their ambiguity.

In our model the structure of the website is obtained from the links between documents and the content is the text extracted from each document. The aim of this model is to generate information that will allow to improve the structure and contents of a website, and also to evaluate the interconnections amongst documents with similar content.

For each query that is submitted to a search engine, a page with results is generated. This page has links to documents that the search engine considers appropriate for the query. By reviewing the brief abstract of each document displayed (which allows the user to decide roughly if a document is a good match for his or her query) the user can choose to visit zero or more documents from the results page. Our model analyzes two different types of queries, that can be found in a website's access registries. These queries are:

External queries: These are queries submitted on Web search engines, from which users selected and visited documents in a particular website. They can be discovered from the log's `referer` field.

Internal queries: These are queries submitted to a website's internal search box. Additionally, external queries that are specified by users for a particular site, will be considered as internal queries for that site. For example, Google.com queries that include `site:example.com` are internal queries for the website `example.com`. In this case we can have queries without clicked results.

Figure 1 (left) shows the description of the model, which gathers information about internal and external queries, navigational patterns and links in the website to discover IS that can be used to improve the site's contents. Also the link and content data from the website is analyzed using clustering of similar documents and connected components. These procedures will be explained in more detail in the following subsections.

3.1 Navigational Model

By analyzing the navigational behaviors of users within a website, during a period of time, the model can classify documents into different types, such as: *documents reached without a search*, *documents reached from internal queries* and *documents reached from external queries*. We define these types of documents as follows:

Documents reached Without a Search (DWS): These are documents that, throughout the course of a session, were reached by browsing and without the interference of a search (in a search engine internal or external to the website). In other words, documents reached from the results page of a search engine and documents attained from those results, are *not* considered in this category. Any document reached from documents visited previously to the use of a search engine will be considered in this category.

Documents reached from Internal Queries (DQ_i): These are documents that, throughout the course of a session, were reached by the user as a direct result of an *internal query*.

Documents reached from External Queries (DQ_e): These are documents that, throughout the course of a session, were reached by the user as a direct result of an *external query*.

For future references we will drop the subscript for DQ_i and DQ_e and will refer to these documents as *DQ*.

It is important to observe that DWS and DQ are *not disjoint sets of documents*, because in one session a document can be reached using a search engine (therefore belonging to DQ) and in a different session it can also be reached without using a search engine. The important issue then, is to register *how many times* each of these different events occur for each document. We will consider the frequency of each event directly proportional to that event's significance for improving a website. The classification of documents into these three categories will be essential in our model for discovering useful information from queries in a website.

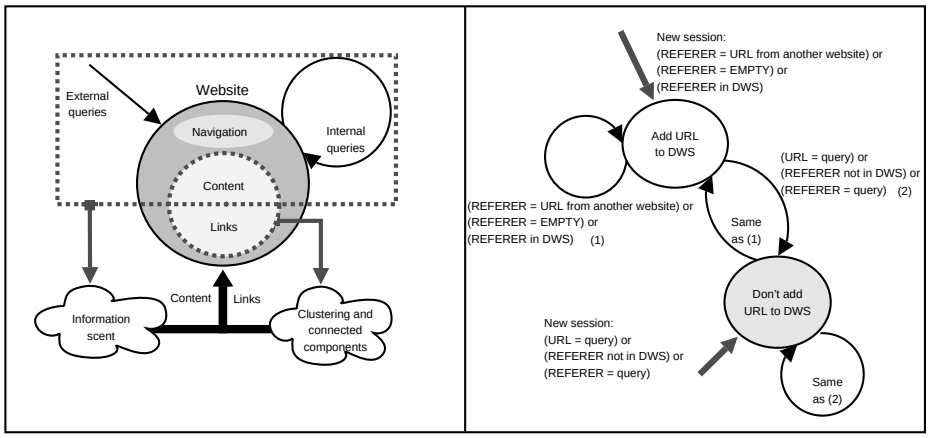

Fig. 1. Model description *(left)* and heuristic for DWS *(right)*.

Heuristic to Classify Documents Documents belonging to DQ sets can be discovered directly by analyzing the referer URL in an HTTP request to see if it is equal to the results page of a search engine (internal or external). In these cases only the *first occurrence* of each requested document in a session is classified. On the other hand, documents in DWS are more difficult to classify, due to the fact that backward and forward navigation in the browser's cached history of previously visited documents is not registered in web servers usage logs. To deal with this issue we created the heuristic shown in figure 1, which is supported by our empirical results. Figure 1 (right) shows a state diagram that starts a new classification at the beginning of each session and then processes sequentially each request from the session made to the website's server. At the beginning of the classification the set DWS is initialized to the value of the

website's start page (or pages) and any document requested from a document in the DWS set, from another website or from an empty referer (the case of bookmarked documents) are added to the DWS set.

3.2 Query Classification

We define different types of queries according to the outcome observed in the user's navigational behavior within the website. In other words, we classify queries according to: if the user choose or not to visit the generated results and if the query had results in the website. Our classification can be divided into two main groups: *successful queries* and *unsuccessful queries*. Successful queries can be found both in internal and external queries, but unsuccessful queries can only be found for internal queries since all external queries in the website's usage logs were successful for that site.

Successful Queries If a query submitted during a session had visited results in that same session, we will consider it as a successful query. There are two types of successful queries, which we will call A and B. We define formally classes A and B queries as follows (see figure 2):

Class A queries: Queries for which the session visited one or more results in *AD*, where AD contains documents found in the DWS set. In other words, the documents in AD have also been reached, in at least one other session, browsing without using a search engine.

Class B queries: Queries for which the session visited one or more results in *BD*, where BD contains documents that are only classified as DQ and not in DWS. In other words documents in BD have *only* been reached using a search in all of the analyzed session.

The purpose of defining these two classes of queries, is that A and B queries *contain keywords that can help describe the documents that were reached as a result of these queries*. In the case of A queries, these keywords can be used in the text that describes links to documents in AD, contributing additional IS for the existing link descriptions to these documents. The case of B queries is even more interesting, because the words used for B queries describe documents in BD better than the current words used in link descriptions to these documents, contributing with new IS for BD documents. Also, the most frequent documents in BD should be considered by the site's administrator as good suggestions of documents that should be reachable from the top levels in the website (this is also true in minor extent for AD documents). That is, we suggest hotlinks based on queries and not on navigation, as is usual. It is important to consider that the same query can co-occur in class A and class B (what cannot co-occur is the same document in AD and BD!), so the relevance associated to each type of query is proportional to its frequency in each one of the classes in relation to the frequency of the document in AD or BD.

Unsuccessful Queries If a query submitted to the internal search engine did not have visited results in the session that generated it, we will consider it as an unsuccessful query. There are two main causes for this behavior:

1. The search engine displayed zero documents in the results page, because there were no appropriate documents for the query in the website.
2. The search engine displayed one or more results, but none of them seemed appropriate from the user's point of view. This can happen when there is poor content or with queries that have polysemic words.

There are four types of unsuccessful queries, which we will call C, C', D and E. We define formally these classes of queries as follows (see figure 2):

Class C queries: Queries for which the internal search engine displayed results, but the user choose not no visit them, probably because there were no appropriate documents for the user's needs at that moment. This can happen for queries that have ambiguous meanings and for which the site has documents that reflect the words used in the query, but not the concept that the user was looking for. Class C queries represent concepts that should be developed in the contents of the website with the meaning that users intended, focused on the keywords of the query.

Class C' queries: Queries for which the internal search engine did not display results. This type of query requires a manual classification by the webmaster of the site. If this classification establishes that the concept represented by the query *exists* in the website, but described with different words, then this is a class C' query. These queries represent words that should be used in the text that describes links and documents that share the same meaning as these queries.

Class D queries: As in class C' queries, the internal search engine did not display results and manual classification is required. However, if in this case, the classification establishes that the concept represented by the query does *not exist* in the website, but we believe that it should appear in the website, then the query is classified as class D. Class D queries represent concepts that should be included in documents in the website, because they represent new topics that are of interest to users of the website.

Class E queries: Queries that are not interesting for the website, as there are no results, but it's not a class C' or class D query, and should be omitted in the classification[4].

Each query class is useful in a different way for improving the website's content and structure. The importance of each query will be considered proportional to that query's frequency in the usage logs, and each type of query is only counted once for every session. Table 1 shows a review of the different classes of queries.

[4] this includes typographical errors in queries, which could be used for a hub page with the right spelling and the most appropriate link to each word.

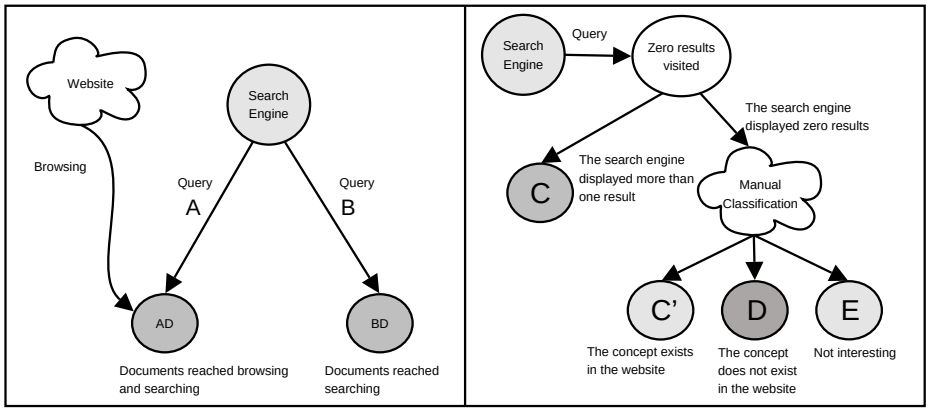

Fig. 2. Successful queries *(right)* and unsuccessful queries *(left)*.

Class	Concept exists	Results displayed	Visited documents	Significance	Contribution	Affected component
A	yes	yes	DQ ∩ DWS	low	additional IS	anchor text
B	yes	yes	DQ \ DWS	high	new IS, add hotlinks	anchor text, links
C	yes	yes	∅	medium	new content	documents
C'	yes	no	—	medium	new IS	anchor text, documents
D	no, but it should	no	—	high	new content	anchor text, documents
E	no	no	—	none	—	—

Table 1. Classes of queries and their contribution to the improvement of a website.

The classification is with memory (that is, an already classified query does not need to be classified in a subsequent usage of the tool) and we can also use a simple thesaurus that relates main keywords with its synonymous. In fact, with time, the tool helps to build an ad-hoc thesaurus for each website.

3.3 Supplementary Tasks

Our Web mining model also performs mining of *frequent query patterns*, *text clustering* and *structure analysis* to complete the information provided by different query classes. We will present a brief overview of these tasks.

Frequent Query Patterns All of the user queries are analyzed to discover frequent item sets (or frequent query patterns). Every keyword in a query is considered as an item. The discovered patterns contribute general information about the most frequent word sets used in queries. The patterns are then compared to the number of results given in each case by the internal search engine, to

indicate if they are answered in the website or not. If the most frequent patterns don't have answers in the website, then it is necessary to review these topics to improve these contents more in depth.

Text Clustering Our mining model clusters the website's documents according to their text similarity (the number of clusters is a parameter to the model). This is done to obtain a simple and global view of the distribution of content amongst documents, viewed as connected components in clusters, and to compare this to the website link organization. This feature is used to find documents with similar text that don't have links between them and that should be linked to improve the structure in the website. This process generates a visual report, that allows the webmaster of the website to evaluate the suggested improvements. At this point, it is important to emphasize that we are not implying that all of the documents with similar text should be linked, nor that this is the only criteria to associate documents, but we consider this a useful tool to evaluate in a simple, yet helpful way, the interconnectivity in websites (specially large ones).

The model additionally correlates the clustering results with the information about query classification. This allows to learn which documents inside each cluster belong to AD and BD sets and the frequency with which these events occur. This supports the idea of adding new groups of documents (topics) of interest to the top level distribution of contents of the website and possibly focusing the website to the most visited clusters, and also gives information on how documents are reached (only browsing or searching).

4 Evaluation

To validate our model we used our prototype on several websites that had an internal search engine, the details of the prototype can be found in [25]. We will present some results from one of those sites: a portal targeted at university students and future applicants. This site has approximately 8,000 documents, 310,000 sessions, 130,000 external and 14,000 internal queries per month. Using our model reports were generated for four months, two months apart from each other. The first two reports were used to evaluate the website without any changes, and show very similar results amongst each other. For the following reports, improvements suggested from the evaluation were incorporated to the site's content and structure. In this approach, the 20 most significant suggestions from the particular areas of: "university admission test" and "new student application", were used. This was done to target an important area in the site and measure the impact of the model's suggestions.

The improvements were made mainly to the top pages of the site, and included adding IS to link descriptions, adding new relevant links, suggestions extracted from frequent query patterns, class A and B queries. Other improvements consisted of broadening the contents on certain topics using class C queries, and adding new contents to the site using class D queries. For example the site was improved to include more admission test examples, admission test scores

and more detailed information on scholarships, because these where issues constantly showing in class C and D queries. To illustrate our results we will show a comparison between the second and third report (we will not show more detailed results due to lack of space). Figures 3, 4, 5 show the changes in the website after applying the suggestions. For figure 5 the queries studied are only the ones that were used for improvements.

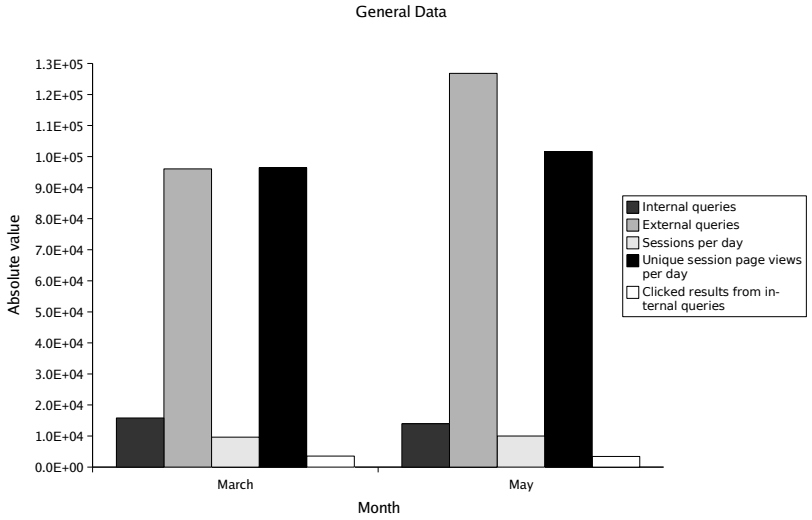

Fig. 3. General results.

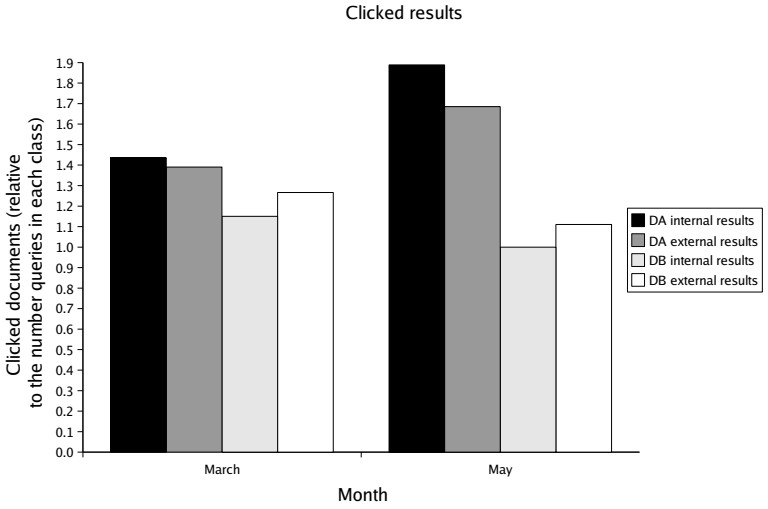

Fig. 4. Clicked results.

In figure 3 we present the variation in the general statistics of the site. After the improvements were made, an important increase in the amount traffic from external search engines is observed (more than 20%), which contributes to an increase in the average number of page views per session per day, and also in the

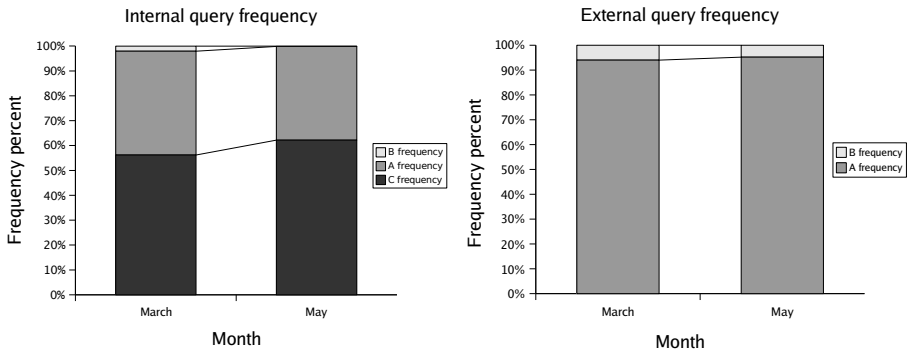

Fig. 5. Internal and External queries.

number of sessions per day. The increase in visits from external search engines is due to the improvements in the contents and link descriptions in the website, validated by the keywords used on external queries. After the improvements were made to the site, we can appreciate a slight decrease in the number of internal queries and clicked documents from those queries. This agrees with our theory that contents are being found more easily in the website and that now less documents are only accessible through the internal search engine. All of these improvements continue to show in the next months of analysis.

Figure 4 shows the comparison between the number of documents (results) clicked from each query class, this number is relative to the numbers of queries in each class. External and internal AD documents present an important increase, showing that more external queries are reaching documents in the website, and that those documents now belong to documents that are being increasingly reached by browsing also. On the other hand BD documents continue to decrease in every report, validating the hypothesis that the suggested improvements cause less documents to be only reached by searching. In figure 5 the distribution of A, B and C queries can be appreciated for internal and external queries. Internal queries (left) show a decrease in the proportion of A and B queries an increase in queries class C. For internal queries (right) A queries have increased and B queries have decreased, as external queries have become more directed at AD documents.

5 Conclusions and Future Work

In this paper we presented the first website mining model that is focused on query classification. The aim of this model is to find better IS, contents and link structure for a website. Our tool discovers, in a very simple and straight forward way, interesting information. For example, class D queries may represent relevant missing topics, products or services in a website. Even if the classification phase can be a drawback at the beginning, in our experience, on the long run it is almost insignificant, as new frequent queries rarely appear. The analysis performed by our model is done offline, and does not interfere with website personalization. The negative impact is very low, as it does not make drastic improvements to the

website. Another advantage is that our model can be applied to almost any type of website, without significant previous requirements, and it can still generate suggestions if there is no internal search engine in the website.

The evaluation of our model shows that the variation in the usage of the website, after the incorporation of a sample of suggestions, is consistent with the theory we have just presented. Even though these suggestions are a small sample, they have made a significant increase in the traffic of the website, which has become permanent in the next few reports. The most relevant results that are concluded from the evaluation are: *an important increase in traffic generated from external search engines*, *a decrease in internal queries*, *also more documents are reached by browsing and by external queries*. Therefore the site has become more findable in the Web and the targeted contents can be reached more easily by users.

Future work involves the development and application of different query ranking algorithms, improving the visualizations of the clustering analysis and extending our model to include the origin of internal queries (from which page the query was issued). Also, adding information from the classification and/or a thesaurus, as well as the anchor text of links, to improve the text clustering phase. Furthermore, we would like to change the clustering algorithm to automatically establish the appropriate number of clusters and do a deeper analysis of most visited clusters. The text clustering phase could possibly be extended to include stemming. Another feature our model will include is an incremental quantification of the evolution of a website and the different query classes. Finally, more evaluation is needed specially in the text clustering area.

References

1. Berendt, B., Spiliopoulou, M.: Analysis of navigation behaviour in web sites integrating multiple information systems. In: VLDB Journal, Vol. 9, No. 1 (special issue on "Databases and the Web"). (2000) 56–75
2. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations **1** (2000) 12–23
3. Cooley, R., Tan, P.N., Srivastava, J.: Discovery of interesting usage patterns from web data. In: WEBKDD. (1999) 163–182
4. Baeza-Yates, R.: Web usage mining in search engines. In: Web Mining: Applications and Techniques, Anthony Scime, editor. Idea Group (2004)
5. Perkowitz, M., Etzioni, O.: Adaptive web sites: an AI challenge. In: IJCAI (1). (1997) 16–23
6. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on web usage mining. Commun. ACM **43** (2000) 142–151
7. Spiliopoulou, M.: Web usage mining for web site evaluation. Commun. ACM **43** (2000) 127–134
8. Batista, P., Silva, M.J.: Mining on-line newspaper web access logs (2002)
9. Cooley, R., Tan, P., Srivastava, J.: Websift: the web site information filter system. In: KDD Workshop on Web Mining, San Diego, CA. Springer-Verlag, in press. (1999)

10. Masseglia, F., Poncelet, P., Teisseire, M.: Using data mining techniques on web access logs to dynamically improve hypertext structure. ACM SigWeb Letters vol. 8, num. 3 (1999) 1–19
11. Huang, Z., Ng, J., Cheung, D., Ng, M., Ching, W.: A cube model for web access sessions and cluster analysis. In: Proc. of WEBKDD 2001 (San Francisco CA, August 2001), 47-57. (2001)
12. Nasraoui, O., Krishnapuram, R.: An evolutionary approach to mining robust multi-resolution web profiles and context sensitive url associations. Intl' Journal of Computational Intelligence and Applications, Vol. 2, No. 3 (2002) 339–348
13. Nasraoui, O., Petenes, C.: Combining web usage mining and fuzzy inference for website personalization. In: Proceedings of the WebKDD workshop. (2003) 37–46
14. Pei, J., Han, J., Mortazavi-asl, B., Zhu, H.: Mining access patterns efficiently from web logs. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. (2000) 396–407
15. Perkowitz, M., Etzioni, O.: Adaptive web sites: automatically synthesizing web pages. In: AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, Menlo Park, CA, USA, American Association for Artificial Intelligence (1998) 727–732
16. Xue, G.R., Zeng, H.J., Chen, Z., Ma, W.Y., Lu, C.J.: Log mining to improve the performance of site search. In: WISEW '02: Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02), Washington, DC, USA, IEEE Computer Society (2002) 238
17. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Ranking boosting based in query clustering. In: Atlantic Web Intelligence Conference, Cancun, Mexico, LNCS Springer (2004)
18. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Web Clustering Workshop at EDBT 2004, Crete, Greece, LNCS Springer (2004)
19. Kang, I.H., Kim, G.: Query type classification for web document retrieval. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM Press (2003) 64–71
20. Sieg, A., Mobasher, B., Lytinen, S., Burke, R.: Using concept hierarchies to enhance user queries in web-based information retrieval. In: IASTED International Conference on Artificial Intelligence and Applications. (2004)
21. Davison, B.D., Deschenes, D.G., Lewanda, D.B.: Finding relevant website queries. In: Poster Proceedings of the Twelfth International World Wide Web Conference, Budapest, Hungary (2003)
22. Baeza-Yates, R.: Excavando la web (mining the web, original in spanish). El profesional de la información (The Information Professional) **13** (2004) 4–10
23. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems **1** (1999) 5–32
24. Mobasher, B.: Web usage mining and personalization. In Singh, M.P., ed.: [**?**]. Chapman Hall & CRC Press, Baton Rouge (2004)
25. Poblete, B.: A web mining model and tool centered in queries. M.sc. in Computer Science, CS Dept., Univ. of Chile (2004)
26. Pirolli, P.: Computational models of information scent-following in a very large browsable text collection. In: CHI. (1997) 3–10