A Content and Structure Website Mining Model

Barbara Poblete
University Pompeu Fabra
Web Research Group
Barcelona, Spain
barbara.poblete@upf.edu

Ricardo Baeza-Yates
University Pompeu Fabra
& CIW, DCC - University of Chile
Barcelona, Spain
ricardo.baeza@upf.edu

ABSTRACT

We present a novel model for validating and improving the content and structure organization of a website. This model studies the website as a graph and evaluates its interconnectivity in relation to the similarity of its documents. The aim of this model is to provide a simple way for improving the overall structure, contents and interconnectivity of a website. This model has been implemented as a prototype and applied to several websites, showing very interesting results. Our model is complementary to other methods of website personalization and improvement.

Categories and Subject Descriptors: H.2.8 [Information Systems]: Data Mining; H.4.m [Information Systems]: Miscellaneous.

General Terms: Performance, Design, Human Factors.

Keywords: Web Mining, Website Improvement.

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 intuitive for its users so they can reach effortlessly the contents they are looking for. Failing to meet this goal can result in the loss of many potential clients.

A website *"is not simply a collection of pages, it is a network of related pages"* [5] and as an interconnected network it can be viewed and studied as a graph. The information provided by the graph organization, along with the contents of the site and the usage data collected by the server, can be very valuable to significantly optimize and enhance a website, thus improving the quality of that site.

To the best of our knowledge, most of the existing website mining models are focused on analyzing the usage information of the website, but they generally do not relate this information with the contents and structure of the site. There is an extensive list of previous work in Web usage mining for improving websites, most of which focuses on supporting adaptive websites [4] and automatic personalization based on Web mining [3]. Amongst other things, using analysis of frequent navigational patterns and association rules [5, 3].

Following this motivation we present a model that uses the existing information in the website, such as its link structure, contents and usage, to help validate the website organization and interconnectivity. The model also suggests, from this point of view, improvements to the website's contents and structure, with the intention of making the contents and link structure more coherent and straightforward for users in general. The main contributions of our model for improving a website are: *to suggest the addition of links between similar documents*, *revise links between unrelated documents*, *point out the most visited sets of documents so they can be linked from the top levels of the website*, and *establish relevant topics in the website based on the most visited clusters of documents, so these topics can be improved in the site.*

2. MODEL DESCRIPTION

Our model, was mentioned for the first time in [1], as part of a larger and more general website mining prototype, centered on user queries. Our model analyzes the contents of a website, based on the text found in each document and the structure of the site, reflected by the links between pages. This information then is processed to validate if the content distribution in the website agrees with its link structure and the usage data registered in the website's access logs. Using this information several reports are generated by the model's prototype, which help the site's administrator visualize possible "problem areas" in the website and ways to solve them.

Text Clustering and Link Analysis

The mining model clusters the documents in the website according to their text similarity (the number of clusters is determined experimentally for each site). In the clustering phase each document in the site is represented internally as a vector of words, in which each coordinate is the frequency of that word in the document. The similarity in this case is measured using the cosine function between vectors, scaled according to the *inverse-document-frequency* paradigm, used in Information Retrieval. It is important to say that the clustering methodology used is a first approximation to this problem, as we are considering substituting this method for a more appropriate algorithm and also obtaining document clusters from the results visited by users from the website's internal (or external) search engines.

The clustering process is achieved using sequential bisections, optimizing in each iteration the global clustering function: $\max(\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(u,v)})$, where k is the total number of clusters, S_i is the number of elements in

WWW 2006, May 22–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

the i-cluster, u, v represent two objects in the cluster and $sim(u, v)$ correspond to the similarity between two objects. This function was experimentally found appropriate for the process, and is discussed in further in [2].

Sequential bisections, generate a hierarchical tree, called a dendogram. The root node of this tree contains all of the documents in the website. Every time a bisection is performed the tree branches into smaller sets until it has as many leafs as desired clusters. The resulting structure is analyzed by our model to validate if documents that are similar, such as documents in the same cluster and documents from clusters that descend from the same parent node, have links between them. We sustain the theory that a user interested in one document is very likely to be interested in other similar documents. In our model we propose that these documents should be connected by links to improve the structure of the website by helping users reach more easily the contents they are looking for.

Correlation of Content and Queries

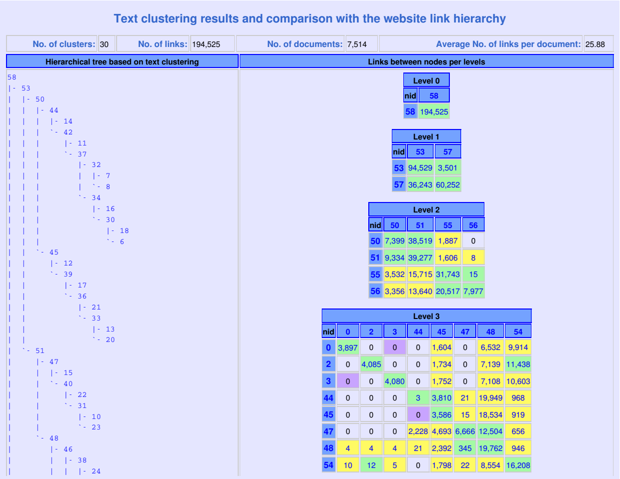
The clustering results are then compared with the information from the usage logs of the website. These logs indicate which clusters were the most visited by users. We propose that the most visited clusters represent "topics" that interest users, and these documents should be linked from the top levels of the website. The usage logs also provide information on which documents were reached by visitors from global search engines and/or from the site's internal search engine. From this we can establish a ratio of documents in a cluster visited from a search engine, compared to the number of documents visited by navigation. This model shows possible problems, such as clusters that are very visited from search engines but not by navigation, which indicates that the links pointing to these clusters are not visible from the top pages of the site or that they do not describe correctly the contents of these documents. In this way the model helps the site's administrator to view these issues and find possible solutions improving the website.

Prototype and Use Case

Our prototype was tested on several sites ranging from small to large, from which we will discuss only one (due to lack of space). The use case belongs to a large portal, targeted towards university students. Our prototype crawled and gathered 7,514 documents, with 194,525 links interconnecting these documents.

The prototype generated a visualization using colors (see figure 1) which shows the dendogram built during the clustering phase and the number of links interconnecting documents inside of the different nodes. Each color represents different levels of interconnectivity amongst documents. The prototype also shows a detailed report with all the usage information regarding each cluster, such as: how many user sessions visited the cluster by navigation, how many user sessions visited the cluster using the internal search engine, or an external search engine, and the most visited clusters in the website. The reports also show the number of connected components inside the cluster, which give a measure of inter-cluster link connectivity.

In the cases in which the prototype has been applied, it has helped the webmasters by to pointing out ways to improve the overall structure of the website: by showing sets of documents that belong to a same topic, but that are not related by links, and learn which topics are more visited, amongst other things. We have found our model specially useful on large sites , in which the contents have become hard to manage for the site's administrator (even for sites that use a content management system).



Text clustering results and comparison with the website link hierarchy

No. of clusters:	30	No. of links:	194,525	No. of documents:	7,514	Average No. of links per document:	25.88

Hierarchical tree based on text clustering

```
58
|- 53
|  |- 50
|  |  |- 44
|  |  |  |- 14
|  |  |  `- 42
|  |  |     |- 11
|  |  |     `- 37
|  |  |        |- 32
|  |  |        |  |- 7
|  |  |        |  `- 8
|  |  |        `- 34
|  |  |           |- 16
|  |  |           `- 30
|  |  |              |- 18
|  |  |              `- 6
|  |  `- 45
|  |     |- 12
|  |     `- 39
|  |        |- 17
|  |        `- 36
|  |           |- 21
|  |           `- 33
|  |              |- 13
|  |              `- 20
|  `- 51
|     |- 47
|     |  |- 15
|     |  `- 40
|     |     |- 22
|     |     `- 31
|     |        |- 10
|     |        `- 23
|     `- 48
|        |- 46
|        |  |- 38
|        |  |  |- 24
```

Links between nodes per levels

Level 0

nid	58
58	194,525

Level 1

nid	53	57
53	94,529	3,501
57	36,243	60,252

Level 2

nid	50	51	55	56
50	7,399	38,519	1,887	0
51	9,334	39,277	1,606	8
55	3,532	15,715	31,743	15
56	3,356	13,640	20,517	7,977

Level 3

nid	0	2	3	44	45	47	48	54
0	3,897	0	0	0	1,604	0	6,532	9,914
2	0	4,085	0	0	1,734	0	7,139	11,438
3	0	0	4,080	0	1,752	0	7,108	10,603
44	0	0	0	3	3,810	21	19,949	968
45	0	0	0	0	3,586	15	18,534	919
47	0	0	0	2,228	4,693	6,666	12,504	656
48	4	4	4	21	2,392	345	19,762	946
54	10	12	5	0	1,798	22	8,554	16,208

Figure 1: Hierarchical tree and interconnectivity.

3. CONCLUSIONS AND FUTURE WORK

In this poster we have presented a new Web mining model that correlates website interconnectivity, contents and overall structure. The aim of this model is to enhance a site, making it more intuitive to its users. Our tool discovers, in a very simple and straightforward way, new and interesting information such as: which topics are more relevant to the website, groups of documents that should be interconnected to each other making easier the navigation in the site, topics that are the most visited from global search engines (these are topics that position the site in the WWW) amongst other useful things. For example, the most visited clusters represent the most relevant topics in the website, and also clusters with a high number of visits from the internal search engine could benefit from better interconnectivity, making it easier for users to reach similar contents to their queries. Our Web mining model is complementary to other methods of website improvement and personalization and we have observed no negative impact after applying the improvements suggested by the model.

Future work involves measuring the improvements due to the suggestions generated by the model, changing our clustering algorithm to automatically determine the best number of clusters for each site, and incorporating clusters of documents that were visited as results of similar queries. Additionally plan to compare our tool with other web traffic analysis systems.

4. REFERENCES

[1] R. Baeza-Yates and B. Poblete. A website mining model centered on user queries. In *EWMF 2005*, pp. 3–15, 2005.

[2] G. Karypis. CLUTO a clustering toolkit. TR. 02-017, Dept. of Cs., Univ. of Minnesota, 2002.

[3] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.

[4] M. Perkowitz and O. Etzioni. Adaptive web sites: an AI challenge. In *IJCAI (1)*, pp. 16–23, 1997.

[5] M. Spiliopoulou. Web usage mining for web site evaluation. *Commun. ACM*, 43(8):127–134, 2000.