Visualization and Collaborative Filtering for Web Mining Tasks

Marko Grobelnik, Dunja Mladenic
Jozef Stefan Institute, Ljubljana, Slovenia
Tel: +386 1 4773900; fax: +386 1 4251038
e-mail: marko.grobelnik@ijs.si

ABSTRACT

The paper proposes several possible tasks for analyzing a web site HTML contents and corresponding Web-Log file using approaches from text visualization and collaborative filtering areas. For this purpose, we used only a simplistic version of the Web-Log file information: a set of Web-Page user accesses in the form of unordered pairs of user- and page-identifications. We developed a recommendation system that enables sharing information about the visits of other users to the same Web page and a system for clustering of the users based on the similarity of their behavior. Additionally, we also cluster the Web pages based on their content and based on the user visits.

1. INTRODUCTION

Web access analysis is an evolving area strongly influenced by the growing number of well-organized Web sites where the owners are interested in improving the site quality and visibility. Each access to the Web site leaves a footprint in the file of the Web site server, containing at least information about the user's computer (IP number) that the request came from and the Web page URL that was requested. Usually we can also use some other fields of information such as: time of the request, URL of the page that the request was made from (for instance following a hyperlink). Sometimes there is additional information obtained by tracking the individual users, for instance by requesting user identification in order to access the Web site. This usually large amount of information is a valuable source of information that is in Web access analysis addressed using Data Mining methods. One of the most popular problems in Web access analysis is finding the most common sequences of Web pages visited one after another, in one session of the user.

In this paper we address the problem of sharing information between the users by identifying the users with similar interests and identifying similar pages. This kind of approach is usually referred to as collaborative approach to user modeling (Mladenic 1998) and can be used for different problems, such as movie or book recommendation (Maes 1994). We extend it on the problem of Web access analysis combining it with clustering of Web pages based on either visits or the page content. We compare the proposed collaborative approach to clustering Web pages to the content-based approach and show directions for combining them. The experiments were performed on the data from Portuguese National Statistics Office having Web site for providing statistical data to the registered users. The Portuguese National Statistics Office (INE) is the governmental agency that is the keeper of national statistical data and has the task of monitoring e.g. country economical data, demographic trends and other important indicators. Their managers believe that the Data Mining technology can give them deeper insights into the understanding of their main web-based online service called "Infoline" (http://www.ine.pt/) – the web site that makes statistical data available to the Portuguese citizens [Jorge and Alves 2001]. Main questions they were interested in were:
(1) How to better understand the structure and the contents of the web site itself?
(2) How to better understand user behavior?
(3) How to reorganize the web site to be more functional?

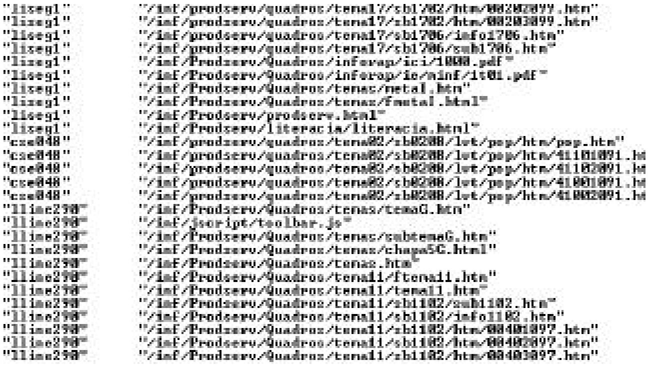
The rest of the paper gives data and tasks definition, with experimental results.

2. DATA DEFINITION

The data we got from the Portuguese National Statistics Office (INE) included (1) documents from the web site and (2) the web-log file with registered user accesses to the web-site documents.

The whole web site contents is in the form of 30.000 mainly HTML and PDF files and their corresponding URLs. The documents are accessible from the web site http://www.ine.pt/.

The original web-server-log files were cleaned and for further experiments we used only approx. 80.000 user accesses to the web-site pages from approx. 1.500 registered users. In the original web-server-log files there were many more records and information but because of how the information about the users was collected, we decided to use just the most reliable part of the data where the identity of the registered user is certain. Here is the sample of the cleaned web-log file – the first column is the user-id (user name for the web-service) and the second column is the corresponding URL of the accessed web page:



```
"liseg1"        "/inf/prodserv/quadros/tema17/sb1702/htm/00202099.htm"
"liseg1"        "/inf/prodserv/quadros/tema17/sb1702/htm/00203099.htm"
"liseg1"        "/inf/prodserv/quadros/tema17/sb1706/info1706.htm"
"liseg1"        "/inf/prodserv/quadros/tema17/sb1706/sub1706.htm"
"liseg1"        "/inf/Prodserv/Quadros/inforap/ici/1000.pdf"
"liseg1"        "/inf/Prodserv/Quadros/inforap/ie/ninf/it01.pdf"
"liseg1"        "/inf/Prodserv/Quadros/temas/metal.htm"
"liseg1"        "/inf/Prodserv/Quadros/temas/Fmetal.html"
"liseg1"        "/inf/Prodserv/prodserv.html"
"liseg1"        "/inf/Prodserv/literacia/literacia.html"
"cse048"        "/inf/prodserv/quadros/tema02/sb0208/lvt/pop/htm/pop.htm"
"cse048"        "/inf/prodserv/quadros/tema02/sb0208/lvt/pop/htm/41101091.h
"cse048"        "/inf/prodserv/quadros/tema02/sb0208/lvt/pop/htm/41102091.h
"cse048"        "/inf/prodserv/quadros/tema02/sb0208/lvt/pop/htm/41001091.h
"cse048"        "/inf/prodserv/quadros/tema02/sb0208/lvt/pop/htm/41002091.h
"llime290"      "/inf/Prodserv/Quadros/temas/temaG.htm"
"llime290"      "/inf/jscript/toolbar.js"
"llime290"      "/inf/Prodserv/Quadros/temas/subtemaG.htm"
"llime290"      "/inf/Prodserv/Quadros/temas/chapa5G.html"
"llime290"      "/inf/Prodserv/Quadros/temas.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/ftema11.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/tema11.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/sb1102/sub1102.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/sb1102/info1102.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/sb1102/htm/00401097.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/sb1102/htm/00402097.htm"
"llime290"      "/inf/Prodserv/Quadros/tema11/sb1102/htm/00403097.htm"
```

3. TASKS DEFINITION

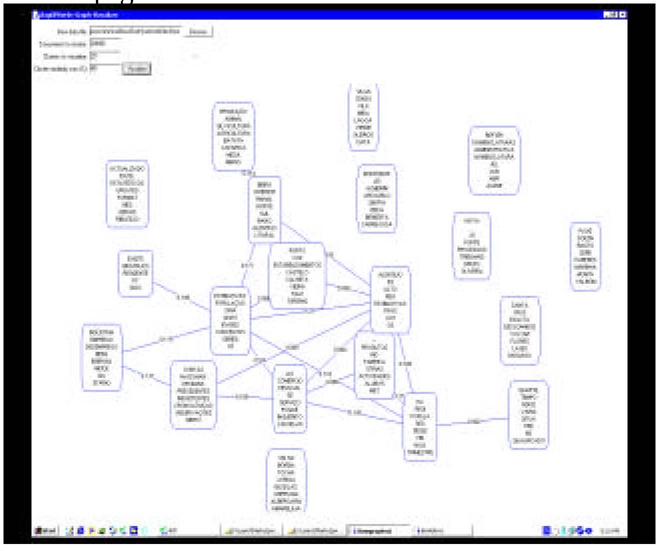
Having the web-site documents and web-accesses usage pairs we defined 4 tasks being interesting for the Portuguese National Statistics Office:

(1) Visualizing the textual contents of the web site.
(2) Page recommendations
(3) Grouping the users according to their behavior.
(4) Visualizing the users behavior.

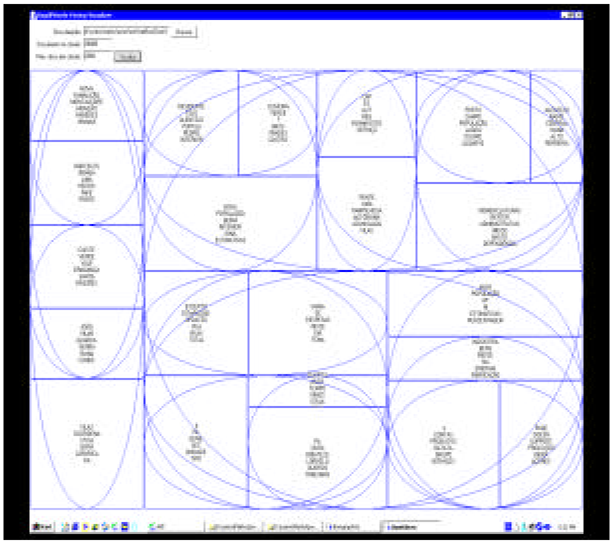
4. WEB-SITE VISUALIZATION

Visualization of the textual contents of a document set is a powerful but not so often used method to get an overview over the contents of documents at the e.g. web-site or some other document collection. In the case of this application we used two methods for text visualization both based on clustering of the documents (see [Manning and Schutze 2001]). The methods are described more in depth in [Grobelnik and Mladenic 2003].

The idea of the first method is: (1) to represent the documents as sparse vectors of real weighted words (using TFIDF weighting scheme), (2) performing K-Means clustering on the whole set of sparse vectors, (3) represent clusters as a graph, where each node in the graph is described by a set of the most characteristic words in the cluster (top weighted words in the centroids of the cluster) and similar nodes (using standards cosine distance) are connected with a link, and finally (4) apply a procedure for nice graph drawing and draw a graph. The resulting picture for all 30.000 pages from the web site is:

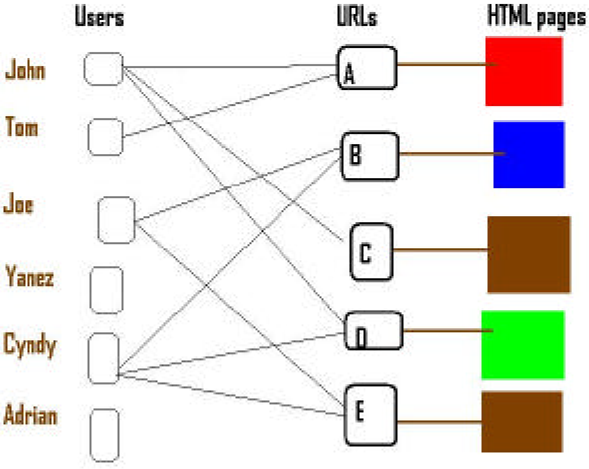


The second method for document visualization again uses sparse vector of TFIDF weights document representation. Instead of standard K-Means it uses hierarchical 2-Means (also called divisive or two-wise clustering), resulting in a tree of clusters having the property that each node consist from a set of documents, its centroid and either being a leaf node or having two child nodes having documents from the parent node split using K-Means clustering, where K=2. Such a tree is visualized using divisive approach starting with the root node owning the whole rectangular drawing area and where it partitions the whole area to the both children proportionally to the number of documents being assigned to them by clustering. This procedure is performed recursively down to the leaf nodes (or if some other stopping criterion applies). The leaf node is represented again by a set of the most characteristic words in the cluster. The resulting picture for all 30.000 documents:



5. WEB-ACCESSESS TASKS

Web access analysis is centered around the collaborative graph formed from the information extracted from the web-log file in the form of [user-id, page-id] set of pairs. The collaborative graph consists from the three main components: (1) nodes, representing the set of all user identifiers (e.g. user names), (2) nodes, representing the set of all page identifiers (e.g. URLs), and (3) the set of text documents corresponding to the URLs. Nodes representing user-ids and nodes representing page-ids form a bipartide graph where the nodes are connected corresponding to the pages a particular user visited. In the example on the following picture we can see a sample collaborative graph showing all three main components and their linkage. Users and URLs are linked according to the user activity. For example, user node [John] and URL-nodes [A], [C] and [D] are linked because the user John clicked on the pages A, B and C and we extracted this information from the web-log file.



Using collaborative graph, we can define the following three tasks: (1) page-ids can be represented by users (which clicked on particular page), (2) users can be represented by a set of page-ids they clicked to, (3) users can be represented by a textual content of pages they clicked to (text they read while browsing the site). In the following three subsections we describe all three approaches.

5.1 Page recommendations

Page recommendations are in the form: "*…users visited this page, visited also the following pages…*". For each target web page we calculate the set of recommended web pages that might be (hopefully) the most interesting for an average user. The recommended web pages must be in certain sense "similar" to the target page. The similarity between pages can be defined in different ways, using different information about the page (e.g. text, layout, usage, etc). In our case we used the assumption that two pages are more similar if they were visited by a similar type of the users. The sketch of the algorithm:

Given: a set of pairs (user, page-id) extracted from the web-log file
Goal: for each page recommend a set of potentially interesting Web pages ("…users visited this page, visited also the following pages…")

Approach:

- Each page is represented by a vector of users (frequency of visits per user)
- Weights of the visits are normalized with TFIDF weighting schema
- Similarity between pages is estimated by cosine distance
- …for each new page generate k-recommended pages using k-nearest-neighbors using the above similarity

5.2 Clustering the users using URLs

Understanding the user behavior is important for every web-site owner. In our case, we identified the characteristic user behaviors by clustering the users using the information from web-log file. For clustering, we must be able to measure "similarity" between the users coming to our web site. In our first approach we assume that the users are more similar if they visit the same URLs. Following this assumption, we represent each user by a vector of visit frequencies of all URLs in the system normalized using TFIDF weighting schema. The similarity between two users is estimated by cosine distance and for clustering we used K-Means clustering. In the example below, we split the users into 10 clusters, each cluster being represented by the most characteristic (most highly weighted) URLs from the centroid. We found clusters of users being interested primarily in e.g. agriculture, demographics etc.:

```
Cluster-0:  [Mean Sim. 0.141] [100 users] - agriculture
'/quadros/inforap/prevagr/minf/1200.pdf':0.394
....
Cluster-1:  [Mean Sim. 0.150] [94 users] – demographic, environment
'/quadros/tema01/sb0101/zip/00200099.exe':0.220
'/quadros/tema01/sb0101/zip/00200099.exe':0.220
'/quadros/tema01/sb0101/zip/00500099.exe':0.156
'/quadros/tema02/sb0201/htm/00703099.htm':0.146
'/quadros/tema01/sb0101/zip/00400099.exe':0.133
'/quadros/tema02/sb0201/htm/00403099.htm':0.129
....
```

5.3 Clustering the users using text they read

In the similar way as in the section 5.2 we can represent the users not by URLs but by the words from the documents behind URLs. The only significant change from the previous approach is that the vectors are represented with frequencies of words from the documents; everything else (TFIDF normalization, cosine distance, K-Means clustering) was the same. The result was a set of clusters describing the groups of users reading the pages with the same content. The intuition in this case is that two users are more similar if they read the documents with the similar contents (documents having more of the same words).

CONCLUSION

We presented 4 approaches to deal with the web-site data of the moderate quality. We presented approaches from the text visualization and collaborative filtering areas applied to the web site and web-usage data. What is lacking in the current work is evaluation of the presented methods.

Acknowledgement

This work was supported by EU project SolEuNet, IST-1999-11495.

References

[Jorge and Alves 2001] Alípio J. and Mário A.A. End of Phase I summary on INE Infoline: a Sol-Eu-Net end-user Phase I project, Sol-Eu-Net IST-1999-11495 Progress Report, 2001.

[Maes 1994] Maes, P., Agents that Reduce Work and Information Overload, Communications of the ACM, Vol. 37, No. 7, pp.30-40, July 1994.

[Manning and Schutze 2001] Manning C.D., Schutze, H. (2001) Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

[Mladenic 1998] Mladenic D., (1998) Text-learning and intelligent agents, IEEE EXPERT, Special Issue on Applications of Intelligent Information Retrieval, 1998.

[Grobelnik and Mladenic 2003] Grobelnik M., Mladenic D. (2003) Two efficient methods for text visualization, Proceedings of TELRI Workshop, 2003