ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Usage Guided Clustering of Web Pages with the Median Self Organizing Map

Fabrice Rossi, Aïcha El Golli and Yves Lechevallier

Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 LE CHESNAY CEDEX – FRANCE

Abstract. Web Usage Mining aims at improving Web sites thanks to the analysis of the behavior of their users. This paper proposes to cluster web pages of a web site thanks to usage data. In big web sites, clustering individual pages is not possible, therefore the proposed method is based on a prior clustering of pages that uses semantic information about the site, such as its organization on the server. Then usage based dissimilarities between prior clusters is defined and an adaptation of the Self Organization Map to such data is used to provide visualization and clustering of groups of pages of the site. The method is illustrated on the web site of INRIA.

1 Introduction

Web Usage Mining (WUM) consists in the analysis of the way a Web site is browsed by its users so as to improve it (in a very broad sense). The practical goals include [1], to name a few: improving the performances of Web servers with intelligent caching and proxy that anticipate user requests; improving the structure of a site based on user typical navigation, for instance by creating automatically site map and bypassing links; introducing recommendations, such as the one proposed by on-line bookstores, based on recognition of typical browsing sessions or of buying patterns.

The goal of the method presented in the present paper is to cluster the content of a Web site based on usage patterns. This allows to understand how users of a Web site perceive it and what kind of relations they build between pages of the site. This will in turn give very valuable information to the web designers: non expected relation between pages might reveal a bad organization of the site (for instance missing links) whereas the absence of strong bound between pages that should be related (on the point of view of the site architects) shows that the content of the pages does not fulfill the need of the *actual* users of the site (who might be very different from the *planed* users).

The paper is organized as follows: in section 2, the data collection process used in WUM is presented. This section also introduces the prior content clustering that is needed to analyze big web sites. Section 3 presents the usage based dissimilarity used to compare prior document clusters as well as the Median Self Organizing Map. Finally, section 4 gives the results of the proposed methodology applied to a real Web site, namely the web site of the institution of the authors.

2 Web Usage Data

2.1 Log files

While some Web sites are only accessible to registered users who must log in to use it and while other Web sites make a systematic use of cookies, many standard web sites use neither of those tricks. This means that the only source of information about user behavior is the log file of the server of the considered web site. This log file consists in the list of all HTTP requests received by a web server with some description of those requests. Interesting information contained in the log file include the IP address of the computer sending the request, the requested document, the date of the request and the *User Agent* that sent the request. This last value describes the web browser used by the user. For instance the user agent `"Mozilla/5.0 (X11; U; Linux i686; rv:1.7.3) Gecko/20041001 Firefox/0.10.1"` corresponds to the Firefox web browser used under a Linux operating system.

The only identification method available in the log file is therefore the IP address and the user agent. Both can be misleading. It is for instance common to have dynamic IP address for home users, or on the contrary to use a Proxy in a corporate environment. If the first case, several IP addresses corresponds to an unique user whereas in the second case an unique IP address can mask several users. The user agent is even more tricky as it can be changed by the user at will. Nevertheless, the use of specific pre-processing algorithms such as [2] allows to reconstruct acceptable (but imperfect) usage data (and also to remove noise coming from automated browsing of a site by indexing robots). If a user is defined by the pair (IP address, User agent), then its trajectory in the web site can be obtained from the log. It consists in a list of requests to documents on the site. The problem of dynamic IP address is reduced thanks to a simple rule: the trajectory of a user is divided into sessions. A session is a list of requests such that the maximum time between two consecutive requests is 30 minutes.

The content clustering proposed in this paper is based only on sessions rather than on user requests. That is, the fact that two sessions might correspond to the same user at different time of the day (for instance), is not taken into account.

2.2 Prior clustering of a web site content

Another difficulty in WUM comes from the fact that sessions are **very** different from each other. On a "big" web site (e.g., more than 100 pages), most of the pages receive a very small number of visits and most of the sessions have nothing in common. Clustering the raw content of a web site, i.e., all its pages, is therefore quite meaningless.

In many WUM algorithms, this problem is solved by focusing on frequent sequences, i.e., on small page sequences that occur frequently in sessions. While this is very useful for some applications such as recommendation systems (e.g., [3]), frequent sequences are not very useful to cluster the content of a web site as they miss most of it (the not so frequent part).

Rather than using frequent sequences, we introduce two major simplifications. First, a prior clustering of the content of the site is performed. The

idea is to define group of documents that are related to each other according to the point of view of the site designers. Indeed, the goal of the proposed method it to put this prior vision under scrutiny and to see whether related documents appears related to the users. This prior clustering can take several forms, depending on the site itself. Many official web sites are well organized with meaningful document names, i.e., Uniform Resource Locators (URL, see [4]). In this case, the simple truncation method proposed in [5] provides a rough clustering that is easy to interpret (it was used in [5] for user clustering). This truncation simply consists in considering that two documents A and B belong to the same cluster if their URLs have the same value up to a given level in the folder hierarchy (an URL is structured into a server name and an unix path, which is in turn broken into a folder hierarchy around the slash character `/`). For instance, documents `http://www-sop.inria.fr/axis/ra.html` and `http://www-sop.inria.fr/axis/` would be in the same cluster if the clustering stops to level one folders (here, `axis`) and in different clusters if it continues to level two folders. The depth of the last considered folder is chosen according to the size of the web site and also according to usage statistics. More precisely, as sessions are in general very different, it is not appropriate to consider thousands of document groups. It is also meaningless to focus on groups that receive a very small number of visits.

Much more complex prior clustering methods can of course be used, for instance content based clustering. Unfortunately, they are in general quite difficult to interpret for the web designers. Indeed, the prior clustering used here corresponds to replacing real documents by URL prefixes. A common best practice in web design is to use meaningful URL prefixes, therefore it is expected that web designers will have no problem to instantaneously understand the meaning of sentences such as "documents whose URL begin with `http://www-sop.inria.fr/axis/`". Nevertheless, the rest of the proposed method can be applied on top of any prior clustering method, even on raw data for small web sites.

The other simplification used in the proposed approach consists in discarding session order, as in [6] for instance. This allows to have a different point of view on the problem. Rather than working on N sessions with different lengths, we now work on K vectors of $\mathbb{N}^N$. K corresponds to the number of prior clusters, i.e., to the number of group of pages. The value x_{kn} is the number of pages from group k that have been requested during session n. Our goal is now to cluster the prior clusters, i.e., the K vectors in $\mathbb{N}^N$.

3 Median Self Organizing Map

3.1 Dissimilarities for WUM

A difficulty induced by the proposed representation is that N (the number of sessions) is in general quite high (more than 16 000 in the proposed application), whereas K is small (around 100 in the application). Moreover, it quickly appears that the euclidean metric in $\mathbb{N}^N$ is completely unadapted to the problem, even when vectors are normalized to avoid size effects (to avoid that long sessions have a bigger weight than short sessions).

Many dissimilarity measure have been proposed to overcome the limitation of the euclidean metric in specific application domains. For user clustering in WUM, the Jaccard similarity index have obtained good results (see e.g., [7]). The dissimilarity based on this index is defined has follows:

$$d(i,j) = \frac{|\{n|(x_{in} > 0 \text{ and } x_{jn} = 0) \text{ or } (x_{in} = 0 \text{ and } x_{jn} > 0)\}|}{|\{n|(x_{in} > 0 \text{ or } x_{jn} > 0)\}|}$$

This dissimilarity is metric but not euclidean [8], which means that the original data cannot be mapped into an euclidean space in which the distance will be equal to the Jaccard index based one. Therefore, traditional vector based methods do not apply to the data considered with this dissimilarity.

3.2 Self Organizing Map for dissimilarities

Kohonen's Self Organizing Map (SOM) [9] fits very well with the considered problem, as it allows to both cluster the data and to give a easy to understand representation of the clusters. Moreover, it has been adapted to dissimilarities with the Median SOM (MSOM) [10]. The main idea of the MSOM (which is a batch SOM) is to adapt the prototype update rule to dissimilarity data. Indeed, the affectation phase consists in finding a winning neuron for each observation. As long as neuron prototypes are chosen among original data, dissimilarities are readily available and the winning neuron is the one whose prototype is the less dissimilar to the considered observation.

The representation phase of the SOM consists in updating the prototype of each neuron so that it represents well the observations affected to it and to its neighbors. In [10], the traditional update rule is replaced by the following problem: find an original observation x_i that minimizes $\sum_{j \in C} d(x_i, x_j)$, where C corresponds to the set of observations associated to the considered neuron and to its neighbors. The problem is solved by a brute force search.

We have used here a variation of the MSOM proposed in [11]. The main differences are that the affectation phase take into account not only one neuron but also its neighbors and that the representation phase uses a weighted sum over observations as in the classical SOM.

4 Application

The proposed approach have been applied to log files coming from three servers of the research institution of the authors (INRIA), `www.inria.fr` which is the main server of the institution, `www-sop.inria.fr` which is the server of the research unit situated in Sophia Antipolis (south of France) and `www-futurs.inria.fr` a newly created research unit (when the log files were recorded). The time period correspond to the first 15 days of 2003. The prior clustering is based on URL truncation after the first folder, i.e., the server name and the first folder are kept. This leads to 107 clusters of URLs that have been visited by at least 5 distinct sessions. There are $N = 16\,717$ sessions for $199\,096$ requests. The URL clusters have been clustered on a 4×3 grid with the MSOM.

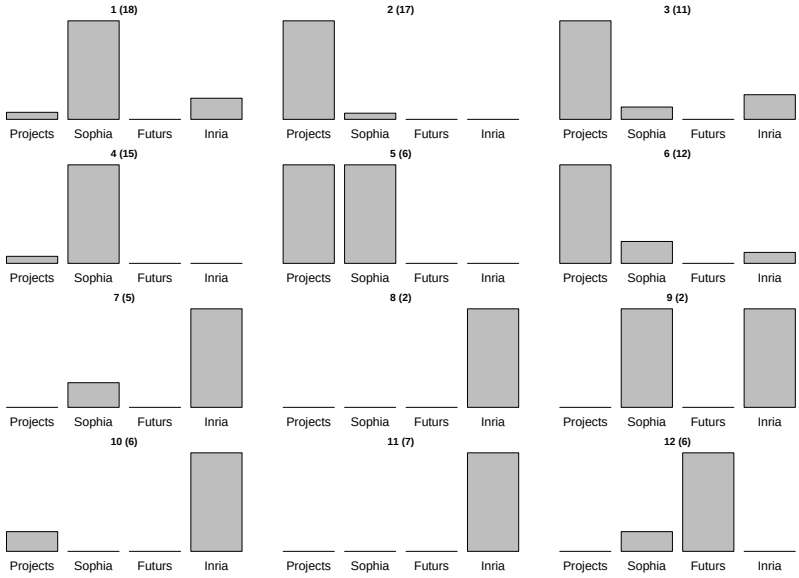

Fig. 1: Clustering results

The final map is given by figure 1. In this figure, each bar graph corresponds to one cluster. The bars display prior information on the prior clusters, more precisely how many different prior clusters of each category has been put in the cluster: the first bar corresponds to prior clusters associated to research team web sites, the other bars to general information on respectively `www-sop.inria.fr`, `www.inria.fr` and `www-futurs.inria.fr`. The first number above each bar graph is the number of the cluster and the second (between parenthesis) is the number of prior clusters affected to this cluster.

This representation shows that clusters are generally uniform regarding to the four categories and spatially organized: the upper left part corresponds to pages from the `www-sop` server, the upper right part of the map contains most pages about research projects, the lower right all pages from the `www-futurs` server, and the lower left is dedicated to pages from the main site server `www.inria.fr`. This good organization means that usage pattern are quite differentiated: users do not make strong links between the main server of the INRIA `www.inria.fr` and pages from the research projects for instance. This might be the sign of design problems in the main site as it contains comprehensive presentation of each research projects together with link to those projects. This remark should be taken with caution as the studied research projects are all part of Sophia Antipolis research unit. Therefore, users interested in such project might browse the web sites starting from the main page of the `www-sop` server, which belongs to cluster 1 and is therefore close to clusters 2, 3, 5 and 6 that contain most of the research project pages.

More precise analysis can be done cluster by cluster. Cluster number 8 for instance contains only the prior clusters `www.inria.fr` (index pages of the

site) and `www.inria.fr/travailler` (pages about recruitment at the INRIA). It appears that 60% of the users that browse pages in the latter prior cluster are coming from the index page of the main site: it seems therefore that the design of this part of the site is correct as it allows users to find what they are looking for (that is, job opportunity at INRIA). Moreover, nearby clusters (5, 7, 9 and 11) contain a lot of prior clusters related to the official communication of the INRIA, such as press releases, industrial contracts, etc. Cluster 11 for instance is mostly dedicated to news about the INRIA.

Cluster 1 contains mostly prior clusters of internal web pages from both the main server and from the `www-sop` server, as well as the root of this server. It seems that users of internal web pages do not type directly the URL of the internal services, but rather browse to them from the main page. Other internal pages can be found in cluster 4.

Overall, the obtained classification is very satisfactory and emphasizes some positive aspects of the web sites (official communication, job opportunities) and some negative ones (a lack of communication between different parts of the web sites).

References

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[2] D. Tanasa and B. Trousse. Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59–65, March-April 2004.

[3] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communication of ACM*, 43(8):142–151, August 2000.

[4] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396, The Internet Society, August 1998. `http://www.ietf.org/rfc/rfc2396.txt`.

[5] Yongjian Fu, Kanwalpreet Sandhu, and Ming-Yi Shih. A generalization-based approach to clustering of web usage sessions. In Masand and Spiliopoulou, editors, *Web Usage Analysis and User Profiling*, volume 1836 of *Lecture Notes in Artificial Intelligence*, pages 21–38. Springer, 2000.

[6] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6(1):61–82, January 2002.

[7] Andrew Foss, Weinan Wang, and Osmar R. Zaïane. A non-parametric approach to web log analysis. In *Proc. of Workshop on Web Mining in First International SIAM Conference on Data Mining (SDM2001)*, pages 41–50, Chicago, IL, April 2001.

[8] J.C. Gower and P. Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48, 1986.

[9] Teuvo Kohonen. *Self-Organizing Maps*. Springer Verlag, New York, 1997.

[10] Teuvo Kohonen and Panu J. Somervuo. Self-organizing maps of symbol strings. *Neurocomputing*, 21:19–30, 1998.

[11] Aïcha El Golli, Brieuc Conan-Guez, and Fabrice Rossi. A self organizing map for dissimilarity data. In D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering, and Data Mining Applications (Proceedings of IFCS 2004)*, pages 61–68, Chicago, Illinois, July 2004. IFCS, Springer.