Improving web site usability through a clustering approach

Martha Koutri[1], *Sophia Daskalaki*[2]

[1]Electrical & Comp. Eng. Dept, HCI Group, [2]Engineering Sciences Dept.
University of Patras, 26500 Rio Patras, Greece
mkoutri@ee.upatras.gr, sdask@upatras.gr

Abstract

One of the main parameters of web usability is related to the smooth navigation of the user and easy access to the information sought. In this paper, we propose a process for improvement of web site structure design by interpreting discovered patterns of web usage logs, with the objective to improve the usability parameters of a given web site. We discuss a cluster-mining algorithm for grouping together related web pages of the site. The clusters found in the process are to be used for redesign of the site and improve connectivity of related pages and thus to facilitate navigation.

1 Introduction

The poor structure of a web site along with the complexity of the information provided quite often result in disorientating its users. On the contrary, user-centered design efforts for web sites almost always lead to the construction of usable web sites, which potentially offer an efficient and agreeable navigational experience. *Usability* is one of the most important features of any software tool and concerns the effectiveness, efficiency, and satisfaction that it gives to the user in a given context of use and task (Tselios, et al., 2001). It is further concerned with making systems safe, easy to learn and easy to use (Preece, 1994). While these facts also hold for web sites, the need for improving the interaction between a site and its visitors, through personalization, is an additional necessity in this domain. Moreover, web personalization requires the development of new methodologies, including web mining techniques. *Web mining* has emerged as a special field during the last few years and refers to the application of knowledge discovery techniques specifically to web data. In addition, the usage patterns, which result by applying web mining techniques to a particular web site log files, can also be used for improving the design of the site. As described in this paper, clustering techniques applied to a web site usage data are especially suitable for discovering those subsets of web documents in the site that need to be connected and those that even if they are connected, their connection should be more visible. Thus clustering results can be used for improving design –and therefore augment the usability– of the web site.
The next section associates the usability of web sites with the browsing behavior of the visitors. Section 3 presents some of the existing approaches for clustering along with our proposed algorithm. Finally, section 4 presents the application of the algorithm to a particular web site in order to demonstrate ways of improving the design of the web site.

2 Exploitation of navigational patterns for improving usability

The application of design guidelines should always direct the design of web interfaces in order for them to be usable. Web design is described using specific terms referring to different areas of concern within the web design space (Newman & Landay, 2000). In particular, the term *information design* refers to the identification of groups of related content and the structuring of information into a coherent whole. *Navigation design* concerns the design of methods for helping users find their way around the information structure. *Graphic design* is the visual communication of information using elements such as color, images, typography, and layout. Many researchers propose supporting web site design from the early phases of the design process. Site maps, storyboards, schematics, etc. are among the tools suggested for representing a web site throughout the design process (Newman and Landay, 2000). It is also very likely that a combination of templates and design conventions will make it easier to design usable web sites (Nielsen, 1999).

Supporting web site design aims at obtaining a desired level of usability. Likewise, in an effort to continuously improve the web site design, even after its implementation, we propose a reconstruction phase where usage patterns revealed from a web usage mining procedure are interpreted to better structures and more usable interfaces. *Web usage mining* concerns the discovery of users' access patterns by analyzing web usage data (Han & Kamber, 2001). *Usage data* is stored into web server access logs and contain information regarding visitors IP address, time and date of access, files or directories accessed, etc. As mentioned earlier, designing the structure of a web site belongs to navigation design. Thus, improvement of the structure of a web site results to more efficient browsing activity, which in fact is related to its usability. Thereby, increased usability is achieved by analyzing navigational behavior of users during their interaction with a given web site.

Besides, even if the design of a web site is user-centered, the mental model of the web site designer differs from users mental models. The concept of mental model describes the cognitive layout that a person uses to organize information on his/her memory (Lokuge, Gilbert & Richards, 1996). Each visitor of a web site has his/her own mental model, which is modulated by his/her needs, desires, cultural and educational level. Thus, each user browses a particular web site by following a specific individual sequence of hyperlinks. The application of web usage mining techniques aims at discovering different navigational patterns, so to facilitate them in a future re-structuring effort of the site.

3 Existing approaches for clustering

Clustering is a knowledge discovery technique based on the idea that similar objects are grouped together and clusters of them are then created. Clustering –unlike classification– is an *unsupervised technique* and this means that there are not any predefined classes and examples of previously observed cases. Standard clustering algorithms partition the set of objects into non-overlapping clusters. Fu, Sandhu & Shih (1999) have presented a system that groups web documents into clusters using the BIRCH algorithm (Zhang et al., 1996). The incremental construction of a CF (Clustering Feature) tree of web documents corresponds with the tree-like graphical representation of web sites. BIRCH is sensitive to the order of entering data records and classifies each document into exactly one cluster. However, the classification of a document into more than one clusters is an important requirement in web mining. Since a web document reserves its own structure, content semantics, and presentation uniformity, different groups of visitors of the site could mentally relate a web document with several different clusters of documents. In (Perkowitz and Etzioni, 2000) a new *cluster mining algorithm*, specifically designed to satisfy the requirements of the web domain, was presented. By learning from visitors' access patterns, they developed PageGather, a cluster mining methodology, for identifying a small set of possibly overlapping clusters, which in fact are collections of mentally related - but currently unlinked - documents in a web site.

The intention of using the results of a cluster mining algorithm to improve navigation design in fact raises the requirement for finding mentally related web documents regardless of their inter-linking with hyperlinks. Suppose, for example, that a cluster consisting of two web documents is the result of a clustering algorithm. This means that a reasonably large number of users tend to visit these two documents in the same session, even if these are not directly linked. The web designer could then improve navigation design of the underlined web site by adding a new bidirectional hyperlink between these two web documents. In another example, consider a cluster consisting of some web documents. Two of these are physically connected via a hyperlink. The classification of both two documents in the same cluster shows their strong mental relation. The designer could therefore highlight the specific hyperlink, so as to help visitors to easily "find their way". In (Avouris et al., 2003) a set of modifications concerning the linkage of web documents and the formatting of hyperlinks are presented. Thus, we ascertain that the application of the appropriate clustering algorithm reveals useful access patterns, which could be used by the web designer in order to improve the usability of a given web site.

3.1 Description of a cluster mining algorithm

The cluster mining algorithm presented here aims at clustering the documents of a web site using information about the presence or absence of each one document during the interaction of different users with the web site. The algorithm takes as input the preprocessed web access logs and generates first all possible singleton clusters from the related web documents. Next, the algorithm successively inserts a second document into the existing clusters to create clusters of two documents. The construction of these two-document clusters is based on the value of a properly defined similarity measure. This measure is a function, which determines a degree of correlation between two or more documents. The process continues by entering more documents to the existing clusters based on the value of the similarity measure, until a desired size for the clusters is reached. The output of the algorithm consists of a set of possibly overlapping clusters of documents that users tend to visit together during their interaction with a web site. We denote that the algorithm has the main characteristics of *agglomerative clustering* approaches (Jain et al., 1999). It begins with each document in a discrete class and proceeds by iteratively inserting documents into the already formed clusters. Moreover, it is insensitive to the order of input data.

In particular, the first step concerns the generation of *user visits* by pre-processing the web access logs. User visits are next used for mining useful information about the usage statistics of web sites. A user visit (or *user session*) is a sequence of page transitions for the same IP address, where each transition is done at a specific time interval (Pierrakos et al., 2001). Thus, each user visit may be represented using a $1\times n$ vector with "0" and "1", where n is the total number of documents in a given web site. A "0" implies that the user has not visited a particular web page, while the value "1" implies that the user has visited it. We next build a $v\times n$ matrix V, where v denotes the number of users' visits available and n the total number of documents in a given web site. The algorithm begins with the formation of singleton clusters by inserting each document into a distinct cluster. Web documents are denoted with the variable i, where $i\in\{1, 2, \dots , n\}$. The algorithm continues by inserting iteratively one document at a time into an existing cluster from previous step according to the value of the *similarity measure* $F_{[i,j,\dots,k]}$, defined as:

$$F_{[i,j,\dots,k]} = \frac{(\#v_i\text{'s in } V\text{, where } p_i = 1, p_j = 1,\dots, p_k = 1)}{(\text{total } \#v_i\text{'s in } V)} \quad (1),$$

where $[i,j,\dots,k]$ is a subset of documents, potentially a cluster, depending on the value of $F_{[i,j,\dots,k]}$, and p_i is a binary variable that denotes the presence ($p_i=1$) or the absence ($p_i=0$) of document i in a certain visit, with $i,j,\dots,k \in\{1,2,\dots,n\}$. If, for example, a singleton cluster from the first step consists of document i, we compute the similarity distance $F_{[i,j]}$, so as to decide if another web document, supposing j, may form a cluster with i in step 2 . The algorithm proceeds by iteratively inserting new web documents into clusters formed in the previous step, until there is not any other insertion to do. The insertion of a web document into an already formed cluster is feasible, if this particular document has not been included yet into the given cluster, and the similarity distance exceeds a predefined threshold *t*. The threshold is an empirically defined parameter by the web miner. Therefore, a group [*i,j,...,k*] of web documents belong to the same cluster *C*, if:

$$(C \ni i) \wedge (C \ni j) \wedge \ldots \wedge (C \ni k) \wedge F_{[i,j,\ldots,k]} \succ t \quad (2)$$

A generalized description of this clustering approach in a Pascal-like pseudocode follows:

```
Input:
Set of n web documents, set of v visits in V, threshold t, size s of each cluster.
Procedure:
    Step 1.0: Form singleton clusters, by inserting each document to a single page clusters.
    Step 2.0: Compute $F_{[i,j]}$, for all $i,j \in \{1,2,\ldots,n\}$ and $i \neq j$.
        Step 2.1: For all $i,j \in \{1,2,\ldots,n\}$ if $F_{[i,j]} \geq t$ then form the 2-page cluster $\{i,j\}$.
    Step 3.0: While $|\{i, j, \ldots, k\}| = s' \leq s$ compute $F_{[i,j,\ldots,k]}$, where $i, j, \ldots, k \in \{1, 2, \ldots n\}$.
        Step 3.1: For all $i,j, \ldots, k \in \{1,2,\ldots,n\}$ if $F_{[i,,j,\ldots,k]} \geq t$ then form the $s'$-page cluster $\{i, j, \ldots, k\}$.
Output:
Set C with overlapping clusters of size s.
```

4 An illustrative example

As an example of the algorithm's applicability, we considered an experimental web site, depicted in Figure 1. After analyzing the web access logs collected from this web site with the cluster algorithm presented above, we received some interesting results regarding its structure.

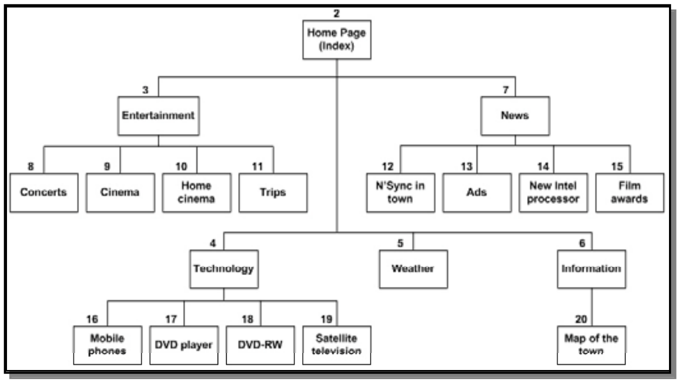

Figure 1. A hierarchical web site

In particular, for threshold *t*=0.30 and clusters size *s*=2, the algorithm returned the following 2-page clusters: {7,3}, {9,3}, and {11, 20}. The web documents "7" and "3" stand for "News" and "Entertainment", respectively. Cluster {7,3} thus indicates that a significant number of users who visit the page with the latest news, also visit the page concerning entertainment. These two documents are not physically linked, because in the designer's mental model were not relate. However, they are related in the users mental model, so, the web designer could add a bidirectional hyperlink conducting directly from the one document to the other. We now consider the cluster {9,3} concerning web pages "Cinema" and "Entertainment", which are physically connected by a hyperlink. Cluster {9,3} indicates that a large number of users who visited page "9", also visited page "3" during the same visit. Regardless of the existence of a hyperlink between the particular documents, an improvement in navigation design would be to make this related information easily accessible. In such a case, adaptation tasks include highlighting, using different colours or fonts, inserting a small icon, etc. Finally, we consider the cluster formed by "11" and "20" representing the web pages "Trips" and "Map of the town". A direct hyperlink could be added, in order to shorten the users navigational paths.

Conclusively, one may say that the clusters of web documents revealed patterns of the users browsing activity. The discovered patterns may then guide the navigation redesign process, in order to improve the usability of a given web site. We note that for different values of threshold we receive different number of clusters. In particular, a medium to large threshold implies a very small number of clusters. In such a case, the web documents, which constitute the corresponding clusters, have large degree of correlation.

5 Conclusions

Clustering the documents in a web site may reveal useful patterns in the browsing activity. The presented innovative cluster-mining algorithm aims at finding groups of related documents of a web site, regardless of their existing links. The presented algorithm is claimed to be more efficient and simpler than other similar approaches. As shown, the resulted clusters can subsequently be used by the designer for improving navigation design of the site, thus, improving its usability.

6 Acknowledgments

Funding by the UoP Research Committee under the basic research program K. Karatheodoris, Project "Development of probabilistic models of web use" is gratefully acknowledged.

7 References

Avouris, N., Koutri, M., & Daskalaki, S. (2003). Web site adaptation: a model-based approach. In Proc. of *HCII2003*, Crete: Greece.

Fu, Y., Sandhu, K., & Shih, M. Y. (1999). Clustering of Web Users Based on Access Patterns. In Proc. of the *1999 KDD Workshop on Web Mining*, Springer-Verlag, San Diego: Canada.

Han, J., & Kamber, M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco.

Jain, A. K., & Dubes, R. C. (1998). Algorithms for Clustering Data. Prentice Hall advanced reference series, Upper Saddle River: NJ.

Lokuge, I., Gilbert, S. A., & Richards, W. (1996). Structuring Information with Mental Models: A Tour of Boston. In Proc. of the *CHI96*, ACM Press, Vancouver, British Columbia.

Newman, M. W., & Landay, J. A. (2000). Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice. In Proc. of the *DIS2000*, ACM Press, New York, 263-274.

Nielsen, J. (1999). User Interface Directions for the Web. *Com. of the ACM*, 42 (1), 65-72.

Perkowitz, M., & Etzioni, O. (2000). Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence 2000* (118), 245-275.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2001). KOINOTITES: A Web Usage Mining Tool for Personalization. In Proc. of *PC HCI 2001*, Patras: Greece, 231-236.

Preece, J. (1994). Human-Computer Interaction. Addison Wesley.

Tselios, N., Avouris, N., Dimitracopoulou, A., Daskalaki, S. (2001). Evaluation of Distance-Learning Environments: Impact of Usability on Student Performance. *Int. J. of Educational Telecommunications*, 7 (4), 355-378.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). An Efficient Data Clustering Method for Very Large Databases. In Proc. of the *SIGMOD*, Montreal: Canada, 103-114.