Web site adaptation: a model-based approach

Nikolaos Avouris[1], Martha Koutri[1], Sophia Daskalaki[2]

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

Abstract

Adaptation of web sites has been an area of great interest during the last years. In this context, development of techniques for knowledge discovery from web usage data have been developed. In this paper a methodological framework is proposed that exploits knowledge discovery from web usage data in support of a process for off-line adaptation of web sites. The proposed approach is based on a web site model consisting of a set of parameters. These parameters define a multidimensional space in which any web site can be located. The adaptation of the web site is based on combination of usage data and the web site model characteristics.

1 Introduction

The development of new knowledge discovery techniques (Fu, Sandhu & Shih, 1999, Mobasher, Cooley & Srivastava, 1999, Perkowitz & Etzioni, 2000) has contributed to the area of adaptation of interaction between users and web sites during the last years. Web site adaptation can take place either on-line, or more often off-line, when usability problems are identified from usage patterns. Our study is focused in off-line web site adaptation that is achieved by applying web usage mining techniques (Srivastava, Cooley, Deshpande & Tan, 2000) and is accomplished by modifying the web site structure. A crucial step in the process is related to the decision on when the adaptation process should be initiated. The process involves *web usage mining*, which concerns the discovery of users' access patterns from web server logs (Han & Kamber, 2001). The adaptation in our case concerns mostly the *web site structure,* i.e. the placement or removal of hyperlinks in the web site's documents. Processing of web usage data constitutes the first important step during the web site structure adaptation process, since from the navigational behavior of users, possible usability problems can be revealed. The majority of methodologies concerning adaptation of web sites involve a knowledge-discovering phase. In addition, an algorithm for suggesting types of adaptation, needs to be incorporated in the process. In this paper, such an algorithm is proposed, as a part of an integrated methodology, which takes into account the specific characteristics of the web site, in order to suggest the most suitable structure. As discussed in the following, the proposed methodology is based on a model of the web site. This model consists of a set of parameters that define the prevailing structure, the character and expected usage of the site without taking into account traffic analysis data.

A fundamental concept in our study is that of the *web site*. According to the W3C (1999) a website is "a collection of interlinked web documents, including a host page, residing at the same network location". However this definition needs further specification, as is demonstrated in the case of off-line browsers, where users have to determine web site boundaries, since there is not a commonly acceptable definition of the boundaries of a web site (Brunk, 1999). A working definition for the purposes of this research identify a web site as *a set of interlinked web* *documents including a host page (which defines the website address), residing in the same network location, that contain links to documents of the same set or to external web documents, cover one or more thematic areas, while they are characterized by thematic coherence, have a uniform presentation, including layout of content and links*. All these defining characteristics guide the selection of the appropriate parameters, in order to construct the web site model, used in our approach.

The next section provides an outline of the proposed adaptation methodological framework. Subsequently, the web site model is defined by presenting its structural elements, while section 4 refers to the possible types of adaptation for the modification of web sites structure, and section 5 discusses the way the model parameters affect the procedure of web site adaptation.

2 An extended methodology for off-line web adaptation

The general procedure for web usage mining consists of the following steps (Cooley, Srivastava & Mobasher, 1997, Mobasher, Cooley & Srivastava, 1999): data collection, data preprocessing, patterns discovery, patterns analysis. This explicit sequence of phases adheres to the fundamental principles of Data Mining. However, the special characteristics of WWW adaptation impose a particular treatment regarding *(i) the knowledge discovery from raw web data, (ii) the application of web usage mining results, in order to adapt the interaction of users with a web site*. In terms of knowledge discovery, many researchers propose variations of existing algorithms or develop new methods and techniques. A generic methodology including a decision support module about web site adaptation needs to be established first. In Figure 1, a proposal for an extended methodology for web usage mining, using a flowchart representation, is shown. The web site modeling module, as well as its interactions with the other processes, constitutes the novel component of the process. In particular, if the analysis of usage patterns -resulted by applying web mining techniques in web log data- identifies the need for adaptation of a particular web site, then the modeling module utilizes the usage patterns, as well as the web site documents structure and the knowledge base, in order to make suggestions about the appropriate types of adaptation. So, a generic methodology that associates web mining and web site modeling processes is structured aiming at supporting adaptation decisions.

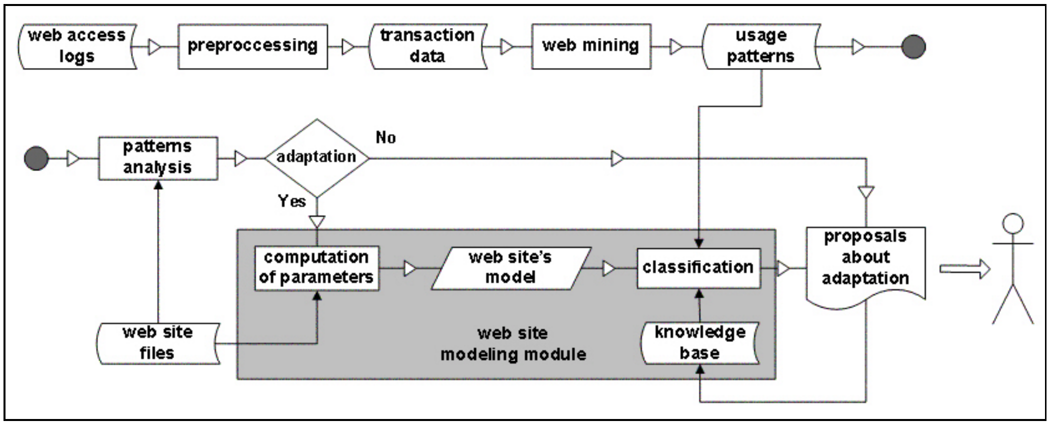

Figure 1. Extended off-line web adaptation methodology

3 A generic web site model

The defined above methodology is based on a web site model. Our attempt to define such a generic web site model in a more formal way yields the definition of such a model as *a descriptive pattern of the web site, which is characterized by the values of some prescribed parameters.* The model parameters need to be defined in such a way that can be related with certain descriptive features of the site. In particular, we propose the following parameters: *address*, *topology*, *type*, *size*, *depth*, *degree of internal connectivity*, and *appearance*. These parameters, relate to the web site working definition, provided in the introduction, and are discussed in the following.

The *Address* concerns the web site's URL specifying the computer that hosts the particular web site and represents its location in the WWW. *Topology* describes the structure of a web site, which results by studying the corresponding web site map. It concerns the order of hyperdocuments in a web site and their inter-connection. *Type* concerns a brief verbal description of the content of a web site. *Size* concerns the number of documents included in the web site. *Depth* measures the number of levels of a given topological structure. *Degree of internal connectivity*, d_{c_in}, is an indicator about the percentage of physically connected web documents in a web site with respect to the total number of documents in the site. In particular: $d_{c\_in}= l_i \cdot 100 / \mathrm{n} \cdot (n\text{-}1)$ (1), where l_i is the number of hyperlinks between the web documents and n is the total number of documents in the site. *Appearance* concerns a brief verbal description about the presentation of a web site.

An empirical study of model applicability and the range of values of the model parameters has been contacted. In particular, we studied the web sites that have been visited by the users of the Greek Research and Technology Network (GRNet), stored in the proxy server of the Network during a period of 6 months (January-June 2002). The range of the topology, type and appearance parameters have resulted from statistical and empirical study of the values of parameters produced from web site modeling. The range for the rest of the parameters has been produced by the study of the sites, and general experience of interaction with the WWW. We note that address is excluded, because it is actually an identifier rather than a descriptive characteristic. Table 1 represents each parameter of the model, a brief definition, as well as the parameter range of values.

Table 1: Parameters of the web site model

No	Parameter	Definition	Range
1	Topology	The structure of the web site.	{sequential, star-like, hierarchical, interlinked, hybrid}
2	Type	Verbal description of the content of the web site.	{business & sales, educational, entertainment, news & media, personal, science & arts, society & politics}
3	Size	Number of hyper-documents in the web site.	{small, medium, large, very large}
4	Depth	Number of levels of the topological structure of the web site.	{small, medium, large}
5	Degree of internal connectivity	$d_{c\_in} = (l_i \cdot 100) / n \cdot (n – 1)$, where l_i is the number of hyperlinks between the web documents and n is the total number of documents in the web site	{small, medium, large, very large, fully-linked}.
6	Appearance	Verbal description of the presentation style of the web site.	{elegant, cute, trendy, rigorous, old-fashioned, ugly, non-artistic}

4 Types of structure adaptation

In this section we present the proposed types of adaptation. A distinction is made between adaptation tasks that concern hyperlink connection and hyperlink formatting. In particular, we describe adaptation of type 1.x concerning hyperlink connection and 2.x for hyperlink formatting.

- *1.1:* Adding new hyperlinks, when some particular web documents are thematically related, while there is not direct physical connection between them. Also bi-directional links, i.e. cross-reference links (Botafogo, Rivlin & Shneiderman, 1992) can be introduced.
- *1.2:* Adding shortcut links, when users visit a particular sequence of web pages, in order to reach a certain target page.
- *1.3:* Removing anchors, when users rarely follow the corresponding hyperlink.
- *1.4:* Totally removing hyperlinks (reference and anchor), when they do not satisfy users needs and desires.

The main types 2.x of adaptation concern formatting of hyperlinks:

- *2.1:* Use of different colors and fonts for highlighting the hyperlink, strongly related with users' needs and desires.
- *2.2:* Use of accompanying icons for recommending or not some particular hyperlinks that interest users or fall outside their scope, respectively.
- *2.3:* Use of a short phrase besides the appropriate hyperlinks for making recommendations - like 2.2 .

As already discussed, the web site model drives the web mining process. This is the subject of the following section.

5 A model-driven adaptation process

The values of the model parameters define the specific web site. The premise of our approach is that decisions of re-structuring the web site should take into account these parameters. In particular, *topology*, d_{c_in}, and *depth* affect connectivity decisions, while *type* and *appearance* affect the formatting of hyperlinks. In addition, we argue that further research needs to be done on the effect that the web site size has on adaptation decisions. So experimentation should be made with various-size web sites, in which evaluation of adaptation decisions should be related to the size of the site. It should be also stressed that the adaptation decisions depend on the combination of parameters for a specific web site. A knowledge base has been constructed after interviews with experienced web-site designers and powerful users. The rules of this knowledge base interrelate the parameters of the website model with adaptation type decisions. An example of this process is given next.

Let us consider an interlinked web site with d_{c_in}=<full linkage>. It has been found that the following 1.x types of adaptation are considered most suitable: 1.2, 1.3, 1.4. Type 1.1 is not considered, because the web site already contains hyperlinks between all its internal documents. In addition, given that depth=<large>, we could concentrate on type 1.2 by adding shortcut links. Besides, if the web site is *educational*, characterized by an *elegant* appearance, then the shortcut links could be highlighted as type 2.1 describes. The application of type 2.2 is not recommended in order to preserve the elegance of the web site. Finally, the educational type of the web site does not lead to application of 2.3 adaptation type. So, we could add a recommendation like "for beginners", in order to help users find their way.

In a similar way, other rules have been derived and a knowledge base has been constructed, which contains a number of possible combinations of values of the model parameters, as well as the corresponding proposals about the types of adaptation. So, the knowledge base contains a set of empirically produced cases, where each one consists of a set of attributes. The attributes are: *topology*, d_{c_in}, *depth*, *appearance* and *type*. Another target attribute, named *recommendation*, is associated to these parameters, containing the recommended types of adaptation for a particular web site. A decision tree corresponding to the given knowledge base was subsequently derived with the use of supervised ID3 algorithm (Quinlan, 1979). The WEKA library of machine learning algorithms in Java (Witten & Frank, 2000) has been used, in order to apply the ID3 classifier into the web site data. This way a generalization of rules that cover other cases, not covered by the experts has been achieved. This knowledge base is combined with usage data for determining the type and place of modification of the web site, as discussed by Koutri & Daskalaki (2003).

6 Conclusions

The described framework has been used in the case of a number of web sites. Evaluation of the effectiveness of this approach is a tedious process, which is still in progress. This involves logging of users' behavior before and after the intervention and collection of subjective user data through questionnaires. A number of modifications in the typology of adaptation types, in the web site model parameters and in the knowledge associating adaptation types to website parameters are expected as a result of this ongoing process. Also various techniques for extraction of patterns of usage need to be tested and used in this process. However one of the main conclusions of the reported research is that off-line web site adaptation needs to be based on models of a generic web site and its structural elements. This model is fully integrated in the web mining process, facilitating decisions and improving traceability of the whole adaptation decision-making process.

7 References

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Trans. on Information Systems* 10 (2), 142-180.

Brunk, D. B. (1999). Overview and Review Tools for Navigating the World Wide Web. SILS Technical Report TR-1999-03.

Cooley, R., Srivastava, J., & Mobasher, B. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. In Proc. of the *9th ICTAI'97*, Newport Beach: Canada.

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. San Francisco: Morgan Kaufmann.

Koutri M., Daskalaki S., (2003). Improving web-site usability through a clustering approach, Proc. HCII2003, June 22-27, Crete.

Mobasher, B., Cooley, R., & Srivastava, J. (1999). Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. In Proc. of the *KDEX'99*. Chicago: Illinois.

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

Quinlan, J. R. (1979). Discovering rules by induction from large collection of examples. Michie, D. (Ed.), Expert System in the Micro Electronic Age, Edinburgh University Press, 168-201.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations* 1: Issue 2.

Witten, I., & Frank, E. (2000). Data Mining: Practical Machine Learning Tools with Java Implementations. San Mateo: Morgan Kaufmann.

W3C (1999) Web Characterization Terminology & Definitions, www.w3.org/1999/05/WCA-terms