Adaptive Interaction with Web Sites: an Overview of Methods and Techniques

Martha Koutri, Sophia Daskalaki, Nikolaos Avouris
University of Patras, 26500 Rio, Patras, Greece
{mkoutri, N.Avouris} @ee.upatras.gr, sdask@upatras.gr

Abstract

The World Wide Web (WWW) tends to become the main medium supporting access to information and services to an expanding group of users. The increasing complexity of existing web applications and the volume of data offered often result in user disorientation. *Adaptive interaction* is proposed by many researchers as an effective means for tackling this problem. In this paper, we present and review existing approaches for developing adaptive web systems. Throughout the review, special emphasis is provided on techniques exploiting web usage data. The survey of methodologies is based on a two-dimensional approach proposed here and used to distinguish different adaptive web systems.

1. Introduction

The World Wide Web (WWW) provides information and services to a continuously growing group of users. Its rapid and continuous expansion contributes to the shaping of our information society. The volume of data in a web site may disorientate the users, who constitute a heterogeneous entirety with different needs and requirements. In this framework, *web personalization* appears as a new application class. Personalization concerns the adaptation of interaction between users and web information space. Mobasher et al. [1] define web personalization as "*any action that tailors the web experience to a particular user, or a set of users*".

Proceedings of the 4th International Workshop on Computer Science and Information Technologies CSIT'2002
Patras, Greece, 2002

This paper attempts a state-of-the-art survey on the adaptation of interaction between users and a web site. Special emphasis is provided on techniques exploiting web usage data. Likewise, the exploitation of the results derived from web usage data analysis is discussed; this expands from an offline transformation of the existing web site to the creation of user profiles and dynamic adaptation.

The rest of this paper is organized as follows: The next section presents the need for adaptive interaction, considering it from two complementary viewpoints. The essential difference between the terms *adaptability* and *adaptivity*, as well as the need for discovering effective methodologies for building adaptive web systems is discussed in section 3. In section 4, we present a state of the art review of existing adaptive systems and approaches. In particular, this concerns on-line recommendation systems including intelligent agents. Next, we present web-log mining as an essential technique for user access patterns discovery and we continue by presenting adaptive web systems and the techniques used for constructing them. Clustering and cluster mining techniques as well as web-log mining are dominant for user profiling. In the following section, directions for clusters exploitation are provided.

2. The need for adaptive interaction

Users interact with a web site in multiple ways, while their mental model about a particular subject is obviously different from these of other users and web developer. Consequently, the improvement of the interaction between users and web sites is of great importance. The adaptation of the desired interaction could be considered from a two-dimensional point of view. The first axis concerns the *adaptation that takes place by making recommendations to the user*, while the appearance and the structure of the web site do not change. The second concerns the *reorganization of the web site and its adaptation to user's needs*. Perkowitz & Etzioni [2] presented *adaptive web sites*: sites that automatically improve their organization and presentation by learning from visitors access patterns.

Adaptive web sites are still a challenge for the researchers of all related scientific fields.

In this paper we focus on the adaptation of interaction between a user and a web site that takes place by modifying a web site. The adaptation is based on *user profiles* or *user models*. User models are defined as models that systems have of users that reside inside a computational environment [3]. User models are completely different from mental models that users have of systems.

3. Rational for automatic adaptation

Before going any further, it is essential to point out the difference between the terms *adaptability* and *adaptivity*. Systems, which allow the modification of certain parameters by the user, are called *adaptable*. On the contrary, systems, which adapt themselves automatically to current user needs or perceived requirements and his/her current task, are called *adaptive*. Adaptable systems include personalized portal sites, like *MyYahoo!* (http://my.yahoo.com), *Go2net* (http://www.go2net.com), *MyExcite* (http://www.excite.com), etc. These sites are supported by systems that guide users to change the background colours, choose content modules or redesign the front- page layout. Every time a user visits the portal, the home page is customized according to his/her preferences. The control of site reformulation is transferred to each user individually, who may change the indicated parameters, having a substantial system support. This manual modification is a useful tool for self-adaptive web interfaces.

However, manual customization provides minimal degree of automation. The user is required to learn the use of the manual adaptation component and therefore the site complexity is further increased. Moreover, with the completion of the "adaptation" task the user finds herself/himself in an improved home page, however still helpless as far as the browsing activity is concerned. The last few years the scientific research has focused on discovering effective methodologies for building adaptive web systems and these are presented next.

4. Web adaptation techniques

At the front of adaptive systems, the research community has proposed a variety of applications for the realization of the desired adaptation. This type of approaches belongs to the new research field of *Web Intelligence* (WI), first introduced by Zhong et al. [4]. In order to facilitate the presentation and review of these applications, we make use of the two- dimensioned approach introduced in section 2, while we take in mind the methods and techniques used. The latter vary and depend on the scientific field, which expands from machine learning and collaborative filtering to probability and graph theory.

4.1 On -line recommendation systems

Contrary to the adaptable systems, a system is adaptive, if it can be adjusted to its users behaviour by learning from their past browsing activity. *Intelligent agents* constitute a particular adaptive systems category. An agent is a software component that assists people and acts on their behalf. For agents, adaptivity is defined as *the ability of the agent to adjust itself to the habits, working methods and preferences of its user*.

Letizia [5] is one of the first intelligent agents, that browses the same space of web documents as the user does and builds a model for each user by observing his/her behaviour. This is accomplished by recording the hyperlinks the user follows, the revisits to documents previously encountered or to a document stored in a hot list. The basic objective of the system is to recommend which of the several possibilities available, concerning the hyperlinks the user could follow, is more likely to satisfy the user.

WebWatcher [6] is another agent that –like *Letizia*– observes user's actions and learns from historical data about the browsing behaviour of the users. Its essential difference from *Letizia* is that the user communicates with the system and gives feedback. The system then processes the information about the current user and her/his interest and the knowledge attained is represented by a utility function. The similarities between the interests of the users and the description of a hyperlink are then computed using reinforcement learning techniques [7]. These are techniques defined as learning what to do e.g. how to map situations to actions, in such a way that a numerical reward is maximized. When the learning procedure has been completed, WebWatcher suggests a number of links that best much the interests of the user by inserting a small icon. Additionally, it provides the capability of search with keywords as well as offering to users commands such as "show me similar pages", "how many users followed this hyperlink", etc.

The systems described above belong to the class of on-line recommendation systems. In [8] a 3-step approach of on-line recommendation systems is presented. According to this approach, the recommendation process includes three phases: (1) the collection of the items to be recommended, (2) the selection of those items from the collected ones that are best suited to the current needs of a particular user and (3) the delivery of the selected items to the user. Both *Letizia* and *WebWatcher* accompany users during their browsing efforts, learn from their actions and make recommendations. Moreover, the learning process is fulfilled by analyzing the actions of the users with the view to infer the user's preferences. In [9] a new type of intelligent agents, known as adaptive web site agents, is

presented. *Adaptive web site agents* are a subclass of intelligent agents. Their main difference with the previous ones concerns the learning process for which web access logs are taken into consideration in addition to the browsing behaviour of the user. A web access log is a complete review of file accesses of a server from the clients and in this case, it is used for the identification of patterns in accessing the site. We discuss this in detail in the next section.

In [10] a newer online recommendation service is presented. In this system the so-called *Broadway* approach is used, where collaborative filtering techniques are applied. Collaborative filtering is an approach that bases automatic adaptation on the opinion of each user for items that are available [11]. The user opinions are collected in a database, which is then used to reveal those users whose opinions present certain similarities and then predict future desires of the users by combining the opinions of like-minded individuals. The *Broadway* system is based on the assumption that if two users follow a similar sequence of web pages, they might have similar browsing intend. Case-Based Reasoning (CBR) techniques have been therefore used to implement this approach, where the sequence of user actions is recorded, analyzed and finally reused.

The system that has been developed in the frame of the *Avanti* project [12] also belongs to on-line recommendation systems. The basic distinction from the other systems described above is that it tries to predict the user's next move. *Avanti* is based on a user model that reflects certain important characteristics of a user, for example his/her background knowledge, preferences, interests and abilities. This kind of information is gained by interviewing users and by recording their dialog actions with the system. When a user requests a web document, the control is transferred to a component of the system, which fetches the requested page from secondary storage. Association rules take into account the user model, and they are then used by the system's inference mechanism. Following this extensive procedure the *Avanti* system is ready to present links leading directly to pages that the system perceives that the user wants to see. The system also highlights additional links that match the user's interests and tries to deduce the user's interests from his/her path through the site.

The *Avanti* system and *WebWatcher* both require the users participation. This means that the user communicates with the system and provides information about herself/himself. Consequently, the system is not transparent, as the user knows of its existence. In general, in the web the nature of interaction necessitates adaptive interaction without users participation for a number of reasons. Systems that act over the shoulder could fully satisfy this demand.

Footprints [13] is a representative system that uses navigation history mining. Specifically, during the navigation of a user on a site, the system records how often each link is traversed. New users may then be guided to use the most frequently traversed paths as indicators of the most interesting pages. Moreover, users need not provide any information about them in order to take advantage of the system. *Footprints* provides a variety of tools such as maps, path views, annotations, comments or signposts by other users, etc..

Footprints is an easy and fast system to use and does not require user identification. However it does not operate dynamically and therefore it does not adapt to user needs and interests. It mainly presents the results of a statistical analysis performed on data representing the history of interaction between users and the web site.

In the following, we will endeavour to study the necessary techniques that need to be applied in order to achieve web adaptation. As already mentioned mining of log data is a necessary step in the process, discussed next.

4.2 Web log mining

The adaptation of human interaction with the web is approached in multiple ways. A common approach, as we have already discussed, is based on the user *browsing behaviour* and focusing on interpreting the way the users browse the web site. For succeeding this goal, it is imperative to analyze large amounts of data collected over the web. Traditional methods from data mining and machine learning need therefore to be adapted for web data.

Web mining has emerged as a special field during the last few years. Web mining is related to data mining and refers to the application of knowledge discovery techniques specifically to web data. Web mining is distinguishing in *web content mining*, *web structure mining* and *web usage mining*, according to the nature of data collected, the objective of each task and to the user needs served. *Web content mining* and *web structure mining* refer to the analysis of content of web pages and structure of links between them, respectively and are beyond the scope of this survey. *Web usage mining*, however, is the process of applying data mining techniques to the discovery of patterns in web data [14]. Web usage mining involves four steps: *user identification, data pre-processing, pattern discovery* and *pattern analysis*.

User access patterns are models of users' browsing activity. In most cases these are deduced from web server access logs. This is referred to as *web-log mining*. A web server access log is a complete review of access of a server from a client. Logs are stored in multiple ways with *Common Log Format* (CLF) to be the most frequently used. Some other formats include *Extended Log Format*

(ELF), *Internet Information Server Format* (IIS) etc.. As an example we note here that a log following the CLF consists of the following elements: client's IP address, user id, access time, request method (get or post), URL, protocol error code, number of bytes transmitted.

Processing of web server access logs is a way for identifying users and is widely used to application of automatic adaptation of a web site. Other alternatives are related to client-side logging, using techniques as cookies. These, as well as the entire procedure of web usage mining are beyond the scope of this paper.

4.3 Clustering based on usage data

Clustering of users is most often the objective of web mining A survey of various clustering approaches is the subject of this section. In general, clustering is a meaningful knowledge discovery technique based on the idea that similar objects are grouped together and clusters of them can be created.

Mobasher et al. [15] proposed a clustering technique using web-log mining to deduce overlapping user preferences profiles. According to this approach, web usage mining takes the form of discovery of overlapping clusters of URLs based on the co-occurrence patterns across user's interaction with the web site. The construction of the clusters consists of two processes, an offline and an online one. The offline process includes the derivation of URL clusters using an association rule mining technique adapted to this type of data. The online process on the other hand refers to making propositions of probably useful links, and is based on user's navigation through the site.

Similarly, Fu et al. [16] have developed a system that groups web users into clusters based on certain common properties. The key idea is the same with that presented in Mobasher et al. [15]. Users behaviour is reflected to web access logs, which are used for the formation of the document clusters. The main difference from the approach mentioned above concerns the techniques used for clusters generation. In particular, Fu et al. propose the construction of a page hierarchy. The hierarchy assumes a partial ordering of the web pages, in which a leaf node represents a web page corresponding to a file in the server (simple page) and a non-leaf node represents a web page corresponding to a directory in the server (general page). Thereby a tree-like structure results. Afterwards, the user's sessions are generalized by replacing the simple page by its corresponding general page. The generalized data are finally clustered using the BIRCH algorithm [17].

While sessions generalization reduces dimensionality of data, it probably could result to the construction of generalized clusters and thus to information loss. The generalized clusters don't include the URLs, but their parent directories.

In the next section some specific clustering algorithms are discussed to address these problems.

4.4 Cluster Mining Algorithms

In general, clustering algorithms take a collection of objects as input and produce a partition of the collection such that each object belongs to exactly one cluster. The classification of a document to a unique cluster, using classic clustering methods is not satisfying for the web domain. In order to find a partition of all web documents we want to find groups of pages that belong together. That means that pages may be placed into more than one cluster. In [18] new clustering algorithms, called *cluster mining algorithms* are presented. These are specifically designed to satisfy the requirements of the web domain. A cluster mining algorithm is an unsupervised algorithm for identifying a small set of possibly overlapping clusters with limited coverage.

PageGather [19] is a cluster mining technique for finding collections of related but currently unlinked documents in a web site. In particular, the algorithm applies web-log mining and cluster mining, producing clusters of related web pages. The frequency of co-occurrence of web pages is computed using probability theory. A *similarity matrix* is then created, that is a quadratic matrix whose cells are filled with the values of co-occurrence frequencies. If two pages are already linked, their corresponding cell is set to zero. The view of the matrix as a graph aims at applying graph algorithms to the task of creating clusters of related pages.

Paliouras et al. [18] presented a modification of *PageGather* algorithm, proposing the normalization of the weights assigned to graph edges. This additional step removes biases for the attributes that appear very often in all users. Additionally, the modified cluster mining algorithm does not restrict its search to clusters of a predefined size k, but it generates all clusters. The cluster finding problem is computationally very complex. However, according to the authors, the implementation of the algorithm is very fast. However, it should be observed that even if the algorithm implementation is effective, clusters consisting of many pages (100 pages for example) bear no practical meaning.

According to [18], the time complexity of *PageGather* is $O(L+N^2)$ where L is the number of page views in the log and N is the number of pages in the web site. The creation of similarity matrix requires O (N^2) and this is the most dominant step, while the cluster mining algorithm presented in [19] has a time complexity greater than or equal to the complexity of PageGather.

4.5 A discussion of adaptive web systems

The presentation of existing applications could be treated in the frame of this paper as a guide for enriching the two-

dimensional approach for adapting the interaction between a user and a web site. Table 1 summarizes the expanded two-fold consideration of web adaptation while it presents the classification of approaches presented here according to the techniques used. Specifically, intelligent and adaptive web site agents belong to online recommendation systems, as well as systems like *Broadway* and *Avanti*. The tools based on navigational history mining, like *Footprints*, also belong to recommendation systems. All the other systems are based to the procedure of site reorganization, even if the existing approaches do not clearly mention the way to accomplish it. The modification of a web site could take place offline or online. It could be fulfilled by exploiting user profiles or by directly interpreting the discovered clusters.

The techniques used by the previously mentioned systems belong to the scientific field of machine learning, but some of these are based on graph and probability theory, as well as reasoning. Intelligent agents are based on user profiling. Particularly, *Letizia* [5] is one of the first attempts to develop an adaptive system based on user modeling. *Letizia*, like *WebWatcher* [6] and *Avanti* [12] for example, attempt to construct users' preferences models by observing their browsing activity. Sometimes a combination of techniques from different fields is used to construct an adaptive system.

A careful study of adaptive systems temporal evolution bears out the tendency of extracting useful knowledge from web access logs. This conclusion results from the fact that most recently developed systems are based on data that characterize the usage of a web site. Web-log mining is a powerful technique for knowledge discovering about users browsing behavior providing the desired transparency.

Clustering and cluster mining techniques equally prevail and shoot mainly at user profiling (see for example [15], [16], [18], [19]). However, existing cluster mining approaches do not provide a clear remark for clusters exploitation. In section 5 we make an attempt for placing the directions for clusters exploitation that essentially leads to effective web site adaptation.

5. Directions for web site adaptation

The clusters produced by applying cluster mining techniques could be used in multiple ways for reorganizing a web site. The simpler is building offline a new modified web site common for all users. *Highlighting* and *linking* are the most common techniques used for making transformations. Highlighting emphasizes to a link by changing its fonts and colors or by adding a small icon. Linking connects two documents not otherwise connected. Web developer has to carefully study the report with the generated clusters for taking a decision about the customization she/he could apply.

		Adaptation of interaction			
		ON-LINE RECOMMENDATION			SITE MODIFICATION
		Agents	Systems	Tools based on navigational history	
Techniques	Association rule mining		AVANTI		Mobasher et al., 1999
	Case-based reasoning		BROADWAY		
	Clustering				Mobasher et al., 1999 Fu et al., 1999
	Cluster mining				PageGather Paliouras et al., 2000
	Collaborative filtering		BROADWAY		
	Graph theory				PageGather Paliouras et al., 2000
	Probability theory				PageGather Paliouras et al., 2000
	Reinforcement learning	WebWatcher			
	Statistical techniques	Letizia, WebWatcher, Pazzani & Billsus, 1999		Footprints	
	Web-log mining	Pazzani & Billsus, 1999	BROADWAY		Mobasher et al., 1999 Fu et al., 1999 PageGather Paliouras et al., 2000

Table 1 Adaptive Web: Systems and Techniques

A more sophisticated solution refers to the formation of user profiles and adaptation based on these profiles. In particular profile creation is accomplished by classifying every user into one or more clusters. Classification is completed using supervised techniques, decision tree learning specifically. Afterwards the web site is adapted in accordance with users' profiles. The adaptation may be static or dynamic. Static adaptation takes place offline. On the contrary, dynamic adaptation eventuates on the fly and provides maximal automation. We underline the necessity for the adaptation of interaction between the user and a web site that takes place during the navigation and is different for each user or group of users.

6. Conclusions

The volume of data in a web site disorientates its visitors who constitute a heterogeneous entirety with different needs and desires. The adaptation of interaction between users and web information space is an important issue. In this paper, we presented and reviewed some existing web systems. We considered a two-dimensional approach for the adaptation of interaction. We then argued that web-log mining is a powerful technique for discovering useful information about the browsing behaviour of the users and most importantly in a transparent way to the system. It is argued that clustering and cluster mining seem to dominate as techniques by measuring the frequency of their use in recently developed systems. We then discussed some ways for the exploitation of the resulted clusters.

This research in the future is expected to focus further on exploitation and experimentation with the proposed

approach with specific kinds of web sites. In this context issues of non intrusive gathering of user information and preferences, as well as on approaches for tackling the sensitive issue of privacy of user data, a key concern of adaptive web, as discussed in [20], are going to be further studied.

Acknowledgements

The research reported here is partially funded by the University of Patras Research Committee under the program of basic research K. Karatheodoris, Project " Development of probabilistic models of web use".

References

1. Mobasher B, Dai H, Luo T, Sun Y, Zhu J. "Integrating Web Usage and Content Mining for More Effective Personalization". In: Proceedings of the *International Conference on E-commerce and Web Technologies* (ECWeb2000), Greenwich, UK, 2000
2. Perkowitz M, Etzioni, O. "Adaptive Web Sites: An AI challenge". In: Proceedings of the *IJCAI-97*, Nagoya, Japan, 1997
3. Fischer G. "User Modeling in Human-Computer Interaction", *User Modeling and User-Adapted Interaction* 2001; 11: 65-86
4. Zhong N, Liu J, Yao Y Y, Ohsuga S. "Web Intelligence (WI)". In: Proceedings of the *24th IEEE Computer Society International Computer Software and Applications Conference* (COMPSAC 2000), 2000, pp. 469-470
5. Lieberman H. "Letizia : An Agent That Assists Web Browsing". In: Proceedings of the *International Joint Conference on Artificial Intelligence*, Montreal, CA, 1995
6. Joachims T, Freitag D, Mitchell T. "WebWatcher: A Tour Guide for the World Wide Web". In Proceedings of the *International Joint Conference on Artificial Intelligence*, 1997
7. Sutton R S, Barto A G. "Reinforcement Learning: An Introduction", MIT Press, Cambridge, MA, 1998
8. Balabanovic M. "An Adaptive Web Page Recommendation Service". In: Proceedings of the *First International Conference on Autonomous Agents* (Agents'97), Marina del Ray, CA, 1997
9. Pazzani M, Billsus D. "Adaptive Web Site Agents". In: Proceedings of the *Third International Conference on Autonomous Agents* (Agents'99), 1999
10. Trousse, B. "Evaluation of the Prediction Capability of a User Behavior Mining Approach for Adaptive Web Sites". In: Proceedings of the *6th RIAO Conference – Content-Based Multimedia Information Access*, Paris, France, 2000
11. Good N, Schafer J, Konstan J, Borchers A, Sarwar B, Herlocker J, Riedl J. "Combining Collaborative Filtering with Personal Agents for Better Recommendations". In: Proceedings of the *Sixteenth National Conference on AI*, 1999
12. Fink J, Kobsa A, Nill A. "User-Oriented Adaptivity and Adaptability in the AVANTI project". In: *Designing for the Web : Empirical*, Microsoft Usability Group, Redmond (WA), 1996
13. Wexelblat A, Maes P. "Footprints: History-rich Web browsing". In: Proceedings of the *Conference on Computer-Assisted Information Retrieval* (RIAO), 1997, pp. 75-84
14. Srivastava J, Cooley R, Deshpande M, Tan P N. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". In: *ACM SIGKDD*, 2000
15. Mobasher B, Cooley R, Srivastava J. "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs". In: *IEEE Knowledge and Data Engineering Workshop* (KDEX'99), 1999
16. Fu Y, Sandhu K, Shih M Y. "Clustering of Web Users Based on Access Patterns". In: Proceedings of the *1999 KDD Workshop on Web Mining*, Springer-Verlag, San Diego, CA, 1999
17. Zhang T, Ramakrishnan R, Livny M. "BIRCH: An Efficient Data Clustering Method for Very Large Databases". In: Proceedings of the *1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, 1996,pp. 103-114
18. Perkowitz M, Etzioni O. "Towards adaptive Web sites: Conceptual framework and case study". *Artificial Intelligence* 2000; 118: 245-275
19. Paliouras G, Papatheodorou C, Karkaletsis V, Spyropoulos C D. "Clustering the Users of Large Web Sites Into Communities". In: ICML 2000, 2000, pp. 719-726
20. Brusilovsky P, Maybury, M.T., From Adaptive hypermedia to the adaptive web, Com. of ACM, 45 (5), pp. 30-33, special issue on the adaptive web, May 2002.