A survey on web usage mining techniques for web-based adaptive hypermedia systems[1]

Martha Koutri, Nikolaos Avouris, Sophia Daskalaki
University of Patras, Greece

Abstract

This chapter discusses web usage mining techniques that can be applied for building adaptive hypermedia systems. These techniques are used for uncovering hidden patterns within web access data and then for building the user model that lies in the heart of each adaptive system. Web access data, traditionally stored in the server log files, constitute a rich source of data collected in a non-intrusive way that guards the privacy of users. Several web usage mining approaches have been proposed for exposing usage patterns with the most prominent ones being cluster mining, association rule mining, and sequential pattern mining. This chapter provides an overview of the state of the art in research of web usage mining, while discusses the most relevant criteria for deciding on the suitability of these techniques for building an adaptive web site. Moreover, the different types of patterns revealed from web usage mining are correlated with different adaptation aspects.

1. Introduction

The web constitutes a huge, widely distributed and dynamically growing hypermedium, supporting access to information and services. Given the diversity of web users, a desired characteristic of many web systems is to personalise user interaction for tasks like navigation and information retrieval. So for instance, it is desirable for e-commerce applications to adapt their offers of products to different customers, for educational systems to present knowledge to different students in accordance to their needs and for information sites to present different items according to the characteristics of the requesting users.

Web-based hypermedia systems that take into consideration certain user characteristics in order to guide user interaction are referred as *adaptive web systems*. The study and development of adaptive web systems constitute an area of active research that stimulates the interest of different research communities, such as user modelling, machine learning, data mining, human-computer interaction, etc. Given the availability of large volume of user data collected over the web and the difficulty to extract user characteristics directly from the users, web usage mining techniques play a vital role in adaptive web systems research and practice (Webb, Pazzani and Billsus, 2001).

The main objective of this chapter is to provide a comprehensive overview of web usage mining techniques used for building adaptive web-based hypermedia systems. Therefore, the main focus is on the main approaches used and their effectiveness in various adaptive web systems. The approaches, discussed in section

[1] in S. Y. Chen and G. D. Magoulas (ed), Adaptable and Adaptive Hypermedia Systems, Idea Publishing Inc., Hershey, 2004

3, are: (a) cluster mining, (b) association rule mining, and (c) sequential pattern mining. These approaches bear different characteristics and can be used in a variety of web systems as discussed in the following.

1.1 Web User Modelling

The *User Model (UM)* constitutes a basic component of every adaptive web system. The UM represents user characteristics that enable the system to distinguish among different users. The UM is built using data that are requested either directly by the users or obtained by logging user interaction. There are different types of user data that can be used for building the UM, for example:

- user's characteristics (such as age, gender, location, etc.);
- user's preferences and interests;
- user's knowledge and skills;
- user's behavioural patterns.

While certain user characteristics may be derived directly from data, others are inferred after processing and interpretation of usage data. For example, preferences and interests of users may be derived after processing behavioural data. User models can refer to either individual users or communities of users. In the latter, each user is considered to be a member of a generic user community and an *aggregate user model* is used for describing the different communities of users.

A number of adaptive systems have been proposed for recommending web documents relevant to users interests (see for example, Armstrong, Freitag, Joachims, and Mitchell, 1995; Lieberman, 1995; Kamba and Sakagami, 1997; Minio and Tasso, 1996; Pazzani and Billsus, 1997). These systems require user participation for building UMs, either by requesting the filling of a form regarding their interests or in other cases by requesting rating of the visited web documents. However, users involvement quite often is considered to be a drawback of such systems. Instead, a non-intrusive transparent collection of user data is considered a preferable approach. This may become effective by automatic collection of access data during each user's browsing task. Given that popular web sites register hundreds of megabytes of information in their web server access log files, the need for specialized techniques that process vast amount of data is well understood. *Data mining*, on the other hand, which refers exactly to the *extraction of knowledge from large amount of data* has proposed many promising techniques applicable in this domain. However, straightforward application of data mining techniques to web user access data present several limitations. In particular,

- Data collected during users navigation are not numeric in nature, while traditional data mining techniques mainly deal with numerical data.
- Noise and data incompleteness are important issues for user access data and there are no straightforward ways to handle them.
- The structure and content of hypermedia systems, as well as additional data, like client-side information, registration data, product-oriented user events, etc., often need to be taken into consideration, necessitating thus the use of specific data integration mechanisms.
- Efficiency and scalability of data mining algorithms is another issue of prime importance when mining access data, because of the very large scale of the problems.
- Conventional statistical measures, like frequency of accessed web documents, are too simple for extracting patterns of browsing behaviour.

It is therefore necessary to device and adapt data mining techniques that tackle some of these issues. Such techniques are discussed in this chapter. In this context, *web usage mining* is the term used for *the application of data mining techniques specifically to raw user access data* with the ultimate goal to reveal patterns of usage. In addition to web usage mining, web mining embraces *structure* and *content mining* methodologies that refer to the analysis of the structure and content of the hypermedia system, respectively.

The remaining of this chapter is organised as following: Section 2 presents the technological and theoretical context of building adaptive web-based hypermedia systems. In section 3, an overview of machine learning techniques used for web usage mining are discussed, while in section 4 the relevance of these techniques to various kinds of adaptive web systems are presented. Lastly, in section 5 the implications of these techniques to future research in the area of adaptive hypermedia are included.

2. Adaptive hypermedia systems – The background

The development of adaptive hypermedia systems evolves in three phases:
(1) collection of user-related data;
(2) processing of user data for building or updating the user model; and
(3) implementation of the user model to provide the appropriate adaptation effect.

Before we proceed with the overview of these three phases, we first provide definitions of some relevant terms as they have been published by the World Wide Web Consortium Web Characterization Activity (W3C WCA) ("WCA", 1999). *User* is a person that interactively retrieve and render resources over WWW with the help of a client software. *Web document* or –equivalently– *web page* is a collection of information spread out over one or more web resources, intended to be rendered simultaneously and identified by a single Uniform Resource Identifier (URI). Thus, URI is the identifier of an abstract or physical resource. The term *user visit* or *server session* describes a collection of user clicks in a single web server during a user session, (Figure 1), where *user session* is a set of user clicks across one or more web servers. Lastly, a *click-stream* is defined as a time-ordered list of *page views*, where a page view is the rendered web document in a specific client application.

The identification of user visits from the user access data is a tedious process performed during the subsequent data processing phase. Various techniques have been proposed for defining sessions from access data, including sessionization heuristics. (e.g. Banerjee and Ghosh 2001, Srivastava et al. 2000). Such a technique is the time-oriented heuristic, where a threshold for the time spent on a specific web document or on the entire hypermedia system during a simple user visit is defined. Pierrakos et al. (2003) provide a survey of the most important methods for user visits identification.

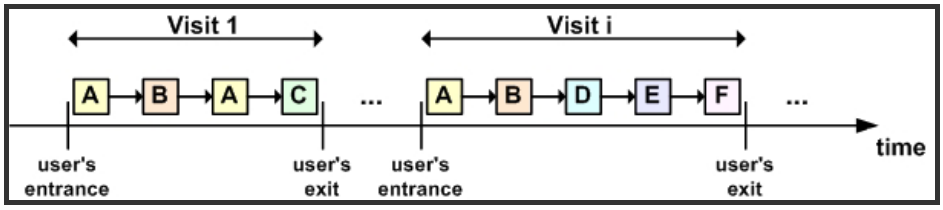

Figure 1. User visits of a particular user in a web server

2.1 Collection of user access data

Web server-side data and client-side data constitute the main sources of data for web usage mining. Pierrakos et al. (2003) present in detail the different data sources and the technological aspects of the techniques used for gathering web data. *Web server access logs* constitute the most widely used data source for performing web usage mining. For this reason, the term *web log mining* is sometimes used. Web log mining should not be confused with *web log analysis*. The former concerns the analysis of server logs by applying data mining techniques, while the latter refers to the traffic analysis performed using mostly statistical techniques. A web server log is a complete review of the access of a specific server from various clients over a period of time. Server logs are stored in a variety of formats depending on the server technology used, such as Common Log Format (CLF), Extended Log Format (ELF), Internet Information Server (IIS) format, etc.

Typically a log entry will include the URL requested, the IP address from which the request originated, a timestamp, as well as additional fields depending on the log format (Table 1).

Field	Value
Client's IP address	213.5.91.177
Client's username *(if a login was required)*	-
Date	6/11/2001
Time	0:10:50
Service *(e.g. WWW, FTP, etc.)*	W3SVC1
Server name	NT-SERVER2
IP address of the server	150.140.184.188
Processing time	2844
Bytes received	386
Bytes sent	22814
Service status code	200
Windows NT status code	0
Name of the operation	GET
Target of the operation	/pchci2001/index.htm
Explanation: An anonymous user accessing the web site from a web client with IP address 213.5.91.177 downloaded (operation GET) the web page /pchci2001/index.htm 10 minutes after midnight at 6 November 2001 from a web server named NT-SERVER2 with IP address 150.140.184.188. The request of the user was processed in 2,844 milliseconds, without errors, and the number of bytes transmitted was 22,814.	

Table 1. The entry of a web server log following the IIS format

User access data carry certain advantages over other types of data provided by the users. Specifically, they are automatically collected "over the user's shoulder", they are of great volume providing a rich data source, they cannot be modified or deleted by the users, and they protect users privacy, as long as the user identity is not provided and cannot be determined.

Limitations encountered with the processing of web access logs refer to the missing document references due to caching mechanisms, the misinterpretation of the IP addresses due to the use of proxy servers assigning the same IP to all users, and the difficulty to update the user model using only access information.

2.2 Processing of user access data

The second phase of the process focuses on the web mining procedures, and involve *data pre-processing*, *pattern discovery*, and *pattern analysis* (Figure 2). We further discuss the pattern discovery sub-process and focus on the nature of the revealed patterns, as well as on the impact of the different types of patterns to the selection of the appropriate adaptation effect. A state-of-the-art overview of all the basic stages of the web usage mining procedure along with the adaptation policies that may be applied to a given hypermedia system can be found in (Pierrakos et al., 2003).

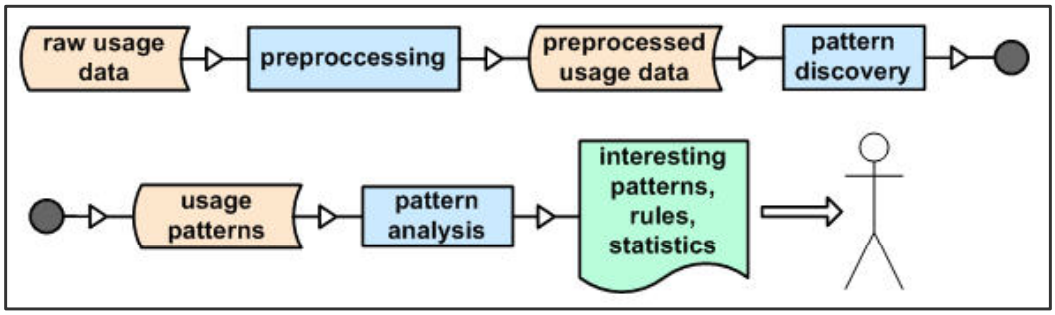

Figure 2. Web usage mining process

Preprocessing refers to the stage of processing the server logs to identify meaningful representations; it consists of *data cleaning* (for removing irrelevant references and fields, removing erroneous references, adding missing references due to caching mechanisms, etc.), *data integration* (for synchronizing data from multiple server logs, integrating registration data, etc.), *data transformation* (for user-session identification, etc.), and *data reduction* (for reducing dimensionality). Data preprocessing can be a difficult task when the available data is incomplete or include erroneous information. A detailed overview of data preprocessing methods can be found in (Cooley, Mobasher and Srivastava, 1999).

The output of the preprocessing phase is used for the *pattern discovery* sub-process. It refers to the actual application of data mining algorithms to the preprocessed data, conducting thus to the discovery of useful patterns. In section 3 of this chapter we discuss some of the most important techniques used for this purpose.

Lastly, *pattern analysis* refers to the interpretation of usage patterns and their presentation in a more attractive way. The most common ways for analysing usage patterns are query-based mechanisms on databases of usage results or On-Line Analytical Processing (OLAP) operations and data cubes (Gray et al., 1997).

2.3 Building the adaptive system

In web-based hypermedia systems adaptation is provided by applying the derived user model during the interaction of the users with the system. While the other two phases of the adaptation process, discussed in sections 2.1 and 2.2, take place offline, this last phase is performed by the system during user interaction.

Conclusively, the overall process of building adaptive systems, following this approach is traditionally supported by two components. The offline component, which refers to data processing for updating or devising the user model, and the online one, which performs the matching of any individual user's browsing activity against the user models in order to provide the appropriate adaptation effect. According to Brusilovsky and Maybury (2002) the adaptive effect can take the form of *content*

selection, *navigation support*, or *presentation*. Thus, the different adaptation effects concern either the automatic selection and prioritisation of the most relevant content items, or the manipulation of hyperlinks like hiding, sorting, annotating, or even the automatic presentation of the content of a given web document. A complete description of the hyperlinks manipulation can be found in (Brusilovsky, 1996). In section 4 of this chapter a discussion on the effect of the web mining approaches to the different types of adaptation is also included.

3. Web usage mining for building adaptive web systems

Web usage mining aims at processing the automatically collected user access data to build suitable user models. *Cluster mining*, *association rule mining*, and *sequential pattern mining* are the most frequently used techniques covering the needs of web usage mining. These are discussed in more detail in the following. It is worth noting that unsupervised learning serves better the needs of web data, because unsupervised learning do not rely on predefined classes or labels for characterizing the data objects. Thus, the difficulty of classifying ambiguous web data using a predefined set of classes is surpassed.

3.1 Cluster mining techniques

Clustering is an *unsupervised classification technique* widely used for web usage mining with main objective to group a given collection of unlabeled objects into meaningful *clusters* (Jain and Dubes, 1998). For the web domain the objects are either *web documents*, or *references of web documents*, or *user visits*. Clustering of web documents is usually based on content data and aims to determine documents with similar content. Therefore, it becomes the subject of web content mining often involving document metadata, document structure data or computational linguistic approaches. On the contrary, in the context of web usage mining, clustering algorithms involve mostly *document references* and *user visits* as input information. For example, let $P = \{P_1, P_2, P_3,..., P_n\}$ be a set of document references corresponding to the n web documents of a given site. Thereupon, clustering of the set P produces clusters of document references and $\{P_1, P_3\}$ could be a potential cluster consisting of two objects. On the other side, for a given web site, we consider a set of user visits $V = \{v_1, v_2, v_3, \ldots, v_m\}$, where each visit is represented as a sequence of web document references, for example $v_1 = (P_1 \rightarrow P_2 \rightarrow P_3 \rightarrow P_{n-1})$. Clustering of the set V derives clusters of visits and $\{v_1, v_2, v_m\}$ could possibly be one of them. In summary, different type of input objects produce different clustering results and thus, impose a variety of usages in the context of adaptive web systems. Specifically, clustering web document references implies *groups of documents* which are intended for similar use, while clustering of user visits reveals *similar navigational patterns* of users. So, the resulting user models represent either different ways of using the hypermedia system or different navigational alternatives that users may adopt.

The application of cluster mining techniques in web systems requires development of new methods for the representation of web objects, because of their non-numerical nature. Some additional factors that influence the performance of clustering algorithms include the viewing time of web documents and the order of the visited documents in a single user transaction. Furthermore, the selection of the appropriate algorithm depends both on the type of data and on the particular purpose for the clustering, since it has great impact on the scalability of the particular

clustering approach. The choice of the function that measures the degree of similarity among web objects affects also the shape of the resulted clusters. For example, algorithms based on Euclidean or Manhattan measures tend to create spherical clusters of similar size. However the formation of clusters with arbitrary shape and size is an important requirement for web usage mining. Lastly, overlapping clusters may or may not be allowed depending on the applications. Traditional clustering uses the following characterizations for the derived clusters:

(1) *exclusive* or *hard* clusters, when an object belongs to one and only one cluster;
(2) *overlapping* clusters, when an object may belong to several clusters, while no additional information is provided about the membership of the objects into the appropriate clusters;
(3) *probabilistic* clusters, where an object belongs to any cluster with a certain probability; and
(4) *fuzzy* clusters, where an object belongs to each cluster with a degree of membership.

In order to review the state-of-the-art for the use of clustering techniques in web usage data a categorization scheme based on the aforementioned factors has been defined, as a guide for comparing the different clustering approaches. In Table 2 the key parameters for our proposed scheme are tabulated.

No	Parameter	Value
1	*Source*	Name(s) of the inventor(s) / Date of publication of the particular approach.
2	*Web objects*	Type of web objects to be clustered (web documents, references of web documents, user visits).
3	*Objects representation*	Method used for representing web objects.
4	*Viewing time*	Yes, if viewing time of web documents is taken into consideration during the representation of web objects. No, otherwise.
5	*Order of click-streams*	Yes, if the order of click-streams is taken into consideration during the representation of web objects. No, otherwise.
6	*Algorithm*	Algorithm used for performing clustering of web objects.
7	*Input data structures*	Input data structures to the clustering algorithm.
8	*Input parameters*	Empirically defined parameters by the data miner.
9	*Similarity measure*	Similarity function used by the clustering algorithm.
10	*Overlapping handling*	Kind of the resulted clusters (hard, overlapping, probabilistic, fuzzy).
11	*Sensitivity to the order of data records*	Sensitive, if the approach is sensitive to the order of data records. Insensitive, otherwise.
12	*Scalability*	Estimate of the scalability of the particular approach (--, -, +, ++).

Table 2. Parameters and definitions

Using this reference scheme, some typical clustering approaches of web usage mining are reviewed next. Tables 3 and 4 provide an overview of those approaches that concern clustering of document references and user visits, respectively.

Clustering of web document produces as output clusters of document references, reflecting thus patterns of common usage. As an example, we consider the cluster $\{A,C,D,E\}$, where A, C, D, E represent web document references. This cluster represents an aggregate user model, which describes that according to current usage of the hypermedia system and regardless of their hyperlink connection, web documents A, C, D, and E are all accessed during a single visit from a set of users with similar characteristics of usage. So, this approach defines clusters of documents that are assumed to be related as they are found often in user visits, while it does not provide any information about the ordering of the accessed documents.

The use of distinctive names for each document reference, as well as the fact that the viewing time and the order of click-streams are not taken into consideration are common characteristics of all approaches in Table 3. Moreover, most of the approaches conduct to the discovery of overlapping clusters, this way capturing common interests among different users or identifying web documents with ambiguous content. We remark that the work of Mobasher et al. (1999) provides a generic framework for the adaptation of a web-based hypermedia system. The other approaches focus on the web usage mining, while the authors propose the use of patterns mined for offline modification of the hypertext system. This observation challenges us to investigate if clustering of web documents gives better results if used for modifying the structure of a hypermedia system in an offline mode rather for adaptive navigation support.

In (Perkowitz and Etzioni, 1997) the *PageGather* algorithm has been proposed for *clustering web document references*. This algorithm learns access patterns by processing the server logs and automatically improves the organization and presentation of the web sites. A further refinement and application of this algorithm is proposed in (Perkowitz and Etzioni, 2000), which addresses the problem of *index page synthesis*. According to this approach, an index page is a page consisting of hyperlinks to a set of pages that cover a particular topic of a given web site. Index page synthesis is the problem of automatically generating index pages to facilitate efficient navigation within the site and to offer a novel view of the site. Perkowitz & Etzioni's approach is performed by clustering the references of web documents and is based on principles of *cluster mining*. Cluster mining algorithms are specifically designed to satisfy the requirement of overlapping clusters and each web document reference is represented by a distinctive name, while the viewing time for each document as well as the ordering of click-streams are not taken into account. The frequency of co-existence of two document references in user visits forms the similarity measure for the cluster mining algorithm. A matrix that indicates the values of all similarity measures between references is then created to guide the formation of a similarity graph. Following that, graph algorithms are applied to identify the *connected components* or the *maximal cliques* and finally to find clusters of web document references. The main drawback of PageGather concerns its low computational performance, which is due to graph algorithms. Such an approach modifies a web hypermedia system semi-automatically. While the formation of clusters is executed automatically in an offline mode, web site modification is manually accomplished by the web master who takes the final decision about what to incorporate in the web site.

In (Mobasher et al., 1999) the focus is on capturing commonalities in the *usage models* with the ultimate goal to perform adaptive navigation support. Here, usage profiles comprise of usage patterns revealed by clustering the references of web documents. The references are represented by unique URLs, for the corresponding

web documents, while viewing times and order of click-streams are not taken into consideration. The overall process is carried out in two modes, an offline and an online. In the offline mode, the algorithm for the generation of *frequent item sets*[2] (Agarwal et al., 1999) is applied to form clusters of document references when two or more documents appear together frequently in user visits. Frequent item sets are then used to form a hypergraph, which is partitioned into a set of clusters using *Association Rule Hypergraph Partitioning* (ARHP). The online component, on the other hand, dynamically provides navigational pointers to users based on their current browsing activity, acting as a real-time recommendation engine.

Source	Algorithm	Input data structures	Input parameters	Similarity measure	Overlap. handling	Sensitivity to the order of records	Scalability
Mobasher et al./1999	Association Rule Hypergraph Partitioning (ARHP)	Similarity hypergraph	–	Connectivity of vertex *v* with respect to a cluster *c*	Overlapping clusters	Insensitive	+
Perkowitz & Etzioni/ 2000	Innovative cluster mining algorithm (PageGather) + Graph algorithms	Similarity graph, data matrix	Graph connectivity threshold, size of clusters, overlapping measure	$\min\{P(p_1\backslash p_2), P(p_2\backslash p_1)\}$	Probabilistic clusters	Insensitive	-
Paliouras et al./2000	Autoclass	Data matrix	Influence threshold	Influence metric	Probabilistic clusters	Insensitive	-
Paliouras et al./2000	Kohonen networks or Self-Organizing Maps (SOM)	Data matrix	Number of clusters, number of time-steps, learning rate, size of neighbourhood function, initialisations of weight vectors, influence threshold	Influence metric	Hard clusters	Insensitive	+ +
Paliouras et al./2000	Variation of PageGather + Graph algorithms	Similarity graph, data matrix	Graph connectivity threshold, clusters overlapping measure	Normalized $\min\{P(p_1\backslash p_2), P(p_2\backslash p_1)\}$	Probabilistic clusters	Insensitive	-

Table 3. Cluster mining of web document references

The problem of assigning the users of a web site into *communities* with common behavioural characteristics was the motivation for applying clustering techniques on document references in (Paliouras et al. 2000). In this case, a distinct name for each document reference has been adopted, and three different algorithms have been applied to provide comparative views of the clusters, named community models. The

[2] A *frequent item set* is a set of items occurring at least as frequently as a pre-defined minimum support threshold.

Autoclass algorithm (Hanson, Stutz, and Cheeseman, 1991), *Self-Organizing Maps* (*SOM*) (Kohonen, 1997), a neural network approach to unsupervised classification, and a *variation of PageGather* the algorithm presented in (Perkowitz and Etzioni, 2000) are thus exploited. Autoclass has a strong mathematical foundation producing clusters of high quality, but is computationally expensive. SOM provides visualization of high dimensional data, but requires a lot of input parameters. Lastly, the variation of PageGather has the main advantages and drawbacks of the original version of the algorithm. Viewing time of documents and order of click-streams again are not used by any of the algorithms. Regarding the type of clusters, the first and third algorithms form overlapping ones, while the second performs hard clustering. This work focuses on the second stage of the overall adaptation process (see section 2.2).

We continue this review by turning to typical approaches of *clustering user visits* (Table 4). Clusters of user visits reveal similarities in the browsing behaviour of users and thus, help to model them. All techniques presented in Table 4 focus on the algorithm used for performing the clustering, as well as on the representation of visits. The way user visits are represented affects the definition of the similarity measure used in each algorithm. While most of the approaches construct hard clusters, the approach in (Nasraoui et al., 2000) considers fuzzy clusters by using a fuzzy mechanism. However, overlapping clusters of user visits could easily become a prerequisite, since any user may very well belong to several clusters based her/his interests and needs. On the other hand, many approaches take into account the viewing time of a web document, as well as the order of click-streams, contrary to the clustering of document references. This is the case because a user visit is not just a collection of documents visited in a given period of time. Instead it may offer information about the order of documents visited and possibly the time spent for each document. Of course the length of viewing time cannot always be considered a reliable measure of users preferences, however it may be included as an indication. After clusters of visits have been formed, they can be used for automatic generation of hyperlinks that guide users based on their perceived browsing intentions.

In (Yan, Jacobsen, Molina, and Dayal, 1996) a methodology for automatic classification of the users of a site according to their access patterns has been proposed. An offline module (*Analog*) performs clustering of user visits, while an online one provides the appropriate adaptation effect by dynamically generating hyperlinks. During the offline processing of web logs, user visits are clustered using *Sequential Leader Clustering* (SLC), which makes use of the Euclidean metric for measuring similarity among user visits. Visits are represented as n-dimensional vectors of document references, while each document is assigned a properly defined weight. Viewing time of web documents may or may not be taken into consideration according to the selection of the weight for representing web documents. Each visit is assigned to only one cluster, thus this approach forms hard clusters. The order of click-streams, however, is not taken into consideration. The offline module is fully implemented, while the online one is briefly described.

In (Shahabi, et al., 1997) an approach for clustering visits aiming at dynamic link generation and pre-fetching of web documents has been proposed. The authors focus on the data processing sub-process. Clustering of user visits is performed with the use of *k-means algorithm* (MacQueen, 1967), an efficient partitioning algorithm that decomposes the data set into a set of k disjoint clusters. User visits are represented as sequences of pairs of documents along with their corresponding viewing time. The authors assert that the sequences of documents in a visit are not enough to properly define a navigation path, since two different hyperlinks could have

identical starting and ending pages. So, they designed and implemented a profiler that captures the links a user has selected, in order to visit a particular web document. As for similarity measure the algorithm uses the *angle* among user visits, while the fundamental aspects of path mining guide the computation of that measure. We denote that their approach takes into account the order of click-streams revealing non-overlapping clusters.

Similarly, a clustering approach for modelling the users of web sites, based on user visits was presented in (Fu, Sandhu, and Shih, 1999). The authors state that finding groups with similar access patterns does not necessarily correspond to page level, particularly when the number of pages in a web site is very large. So, they defined a page hierarchy in a tree-like representation, where a leaf node represents a page corresponding to a file in the server, and a non-leaf node represents a page corresponding to a directory. The former is called *simple page* and the latter *general page*. Consequently, user visits are represented as generalized vectors in the form of (session-id, t_1, t_2…, t_n), where t_i is the total time the user spent on the i-th general page and the order of web documents visited is not taken into account. For the clustering procedure the authors used *BIRCH* (Zhang, Ramakrishnan, and Livny, 1996), which requires a *Clustering Feature tree* as input, as well as a branching factor and a diameter threshold as input parameters. For measuring similarity among different visits, Euclidean distance is used. Like with previous approaches, this methodology also produces non-overlapping clusters. Again in this approach, the focus is on the web usage mining sub-process while the use of clustering output is left to the web master.

Along the same lines, in (Nasraoui et al., 2000) the problem of clustering user visits is studied in order to extract typical user session profiles. The main difference from previously mentioned approaches is the use of a fuzzy mechanism, which results to fuzzy clusters. In addition, the representation of visits comprise viewing time for each page visited and the order of click-streams. The clustering methodology here is based on the *Relational Fuzzy C-Maximal Density Estimator* (RFC-MDE). Lastly, a new similarity measure capturing both individual documents in a cluster as well as the structure of the site has been proposed.

In (Banerjee and Ghosh, 2001) user visits are clustered based solely on click-streams. The *Metis* algorithm (Karypis and Kumar, 1998) has been used for clustering user visits, which are represented as sequences of ordered pairs of web pages visited and corresponding viewing times. A new similarity measure consisting of two parts, *similarity* and *importance components* has also been proposed. The Metis algorithm takes as input a similarity graph, in order to reveal the desired clusters. Unlike the fuzzy approach, presented earlier, the clusters discovered by the Metis algorithm are exclusive.

Lastly, the clustering algorithm *TURN* (Foss, Wang, and Zaiane, 2001), has been adopted for grouping user visits by Wang and Zaiane (2002). Users visits are represented as sequences of URLs and the similarity between sequences is computed by using a new alignment measure based on dynamic programming. A similarity matrix is then created and is used as input to TURN. Concerning applicability, this approach is proposed in the context of e-learning to determine learning behaviour of students.

Source	Objects represe-ntation	Viewing time	Order of click-streams	Algorithm	Input data structures	Input parameters	Similarity measure	Overlap. handling	Sensitivity to the order of records	Scalability
Yan et al./1996	*n*-dim vectors of document references	Yes/No	No	Sequential Leader Clustering (SLC)	Data matrix	Radius *r* around a cluster's center, min # of documents in a visit, max value of similarity measure	Euclidean metric	Hard clusters	Sensitive	+
Shahabi et al./1997	Pairs of document-time sequences	Yes	Yes	k-means	Path angle matrix	# of clusters *k*, threshold angle θ_{Th}	Angle among user visits (path mining)	Hard clusters	Insensitive	+ +
Fu et al./1999	Generalized vectors of time sequences	Yes	No	BIRCH	Clustering Feature (CF) Tree	Branching factor *B*, diameter threshold *T*	Euclidean metric	Hard clusters	Sensitive	+
Nasraoui et al./1999	n-dim vectors of appearance (1) or not (0) of a document in a session	No	No	Relational Fuzzy *C*-Maximal Density Estimator (RFC-MDE)	Relation matrix	# *C* of rows from the relation matrix, max # of operations (optional)	Novel metric capturing both the documents in a cluster and the structure of the hypermedia	Fuzzy clusters	Insensitive	+
Banerjee and Ghosh/ 2001	Pairs of document-time sequences	Yes	Yes	Metis	Similarity graph	Edge threshold θ, imbalance threshold *t*	Novel metric consisting of *similarity* and *importance* components	Hard clusters	Insensitive	+
Wang and Zaiane/ 2002	Document sequences	No	Yes	TURN	Similarity matrix	–	Novel metric based on dynamic programming	Hard clusters	Insensitive	+

Table 4. Cluster mining of user visits

3.2 Association rule mining

Association rule mining refers to the identification of all associations among certain data items, so that the presence of one subset of them in a transaction implies the presence of others also (Mobasher, et al. 1996).

The problem of mining a large collection of data for finding association rules between different items was defined in (Agrawal, Imielinski, and Swami, 1993) and an efficient algorithm for generating all significant association rules, known as *Apriori* algorithm, has been presented in (Agrawal and Srikant, 1994). Association rule mining has drawn a lot of attention lately and a variety of algorithms have been developed (e.g. Hipp, Guntzer, and Nakhaeizadeh, 2000).

An association rule is a statement of the form $A \Rightarrow B$, where A and B are disjoint subsets of a set of items. The rule is accompanied by two meaningful measures, *confidence* and *support*. *Confidence* measures the percentage of transactions containing A that also contain B (i.e. *confidence*$(A \Rightarrow B) = P(B \mid A)$). Similarly, *support* measures the percentage of transactions that contain A or B (*support*$(A \Rightarrow B)=P(A \cup B)$) (Han and Kamber, 2001). Association rule mining of web data, on the other hand, describes techniques that are specifically designed for uncovering associations among web documents usually visited by users in a single visit to a particular web site. The resulting associations provide useful knowledge about usage patterns from users, which may further be used for the adaptation process. For example, the association,

e-class/asp_fundamentals.html $\Rightarrow$ *e_class/asp_examples.html*,

with *support*=5% and *confidence*=70% reveals that visitors of a given e-learning web site who access the web page about the fundamentals of ASP, also tend to access the web page referring to some ASP examples. In addition, 5% of all user visits concern visits that include the web pages about ASP fundamentals or ASP examples, while 70% of the users who visited the *e-class/asp_fundamentals.html* page also visited the *e_class/asp_examples.html* web page. This type of rule does not provide any information about the order of pages accessed during a particular visit, however it captures certain relationships among document references.

In this section a number of association rule mining techniques used for web data is presented and are compared in the basis of a set of parameters displayed in Table 5.

No	Parameter	Value
1	***Source***	Name(s) of the inventor(s)/Date of publication of the particular approach.
2	***Graph structure***	Yes, if the graph structure of data is taken into consideration.
3	***Input data structures***	Input data structures to the sequential pattern mining algorithm.
4	***Input parameters***	Empirically defined parameters by the data miner.
5	***Algorithm***	Algorithm used for mining sequential patterns.
6	***Type of patterns***	The type of patterns mined by applying the specific algorithm.

Table 5. Association Rule Mining approaches scheme

(Mobasher et al., 1996) present a web usage mining system, called *WEBMINER*. Its main purpose is the revealing of usage patterns in a given web site, based on the application of several data mining techniques. First, the clustering of user access log entries aims at grouping together user transactions based only on time. Next,

traditional algorithms for mining association rules from the transactions database are applied by determining the value for the support threshold. In a follow-up work, Mobasher, Berendt, and Spiliopoulou (2001) propose the generation of recommendations for the adaptation process directly from the frequent item sets, thus avoiding the generation of all association rules.

Borges and Levene (1998) have proposed the extraction of composite association rules from the structured data of WWW. In this work the notion of confidence and support measures are formalized in the context of directed graphs and two algorithms are proposed. The first is a modification of the Depth-First-Search algorithm and the other uses an incremental approach for mining association rules.

Another methodology for mining fuzzy association rules using a case-based reasoning approach is proposed by Wong, Shiu, and Pal (2001). An important issue here is the definition of cases and the authors propose a schema based on the length of the user transaction. It is claimed that mining fuzzy correlations among web document references leads to more accurate predictions of user access paths.

Table 6 summarizes the key points of the approaches reviewed of this section, using the parameters defined in Table 5.

Source	Graph structure	Input data structures	Input parameters	Algorithm
Mobasher et al/1996	No	Database of user transactions	User-specified maximum time gap, support threshold	Traditional clustering, association rule mining algorithms
Borges & Levene/ 1998	Yes	Database of user transactions, directed graph of the hypermedia	Support, confidence thresholds	Modified Depth-First-Search algorithm, an incremental step algorithm
Wong et al/ 2001	No	Database of user transactions	Support, confidence thresholds	Traditional fuzzy association rule algorithms

Table 6. Association rule mining approaches

Comparing association rule mining with clustering of document references, we perceive that both methods aim at identifying patterns regarding the usage of a set of web documents, while none of them offers any information about the order of documents visited. Clustering and association rule mining apparently provide the same type of results. However, there is an important difference. An association rule provides additional information about the antecedent and the consequent of a pattern, as well as about the values of confidence and support measures.

The above observations could motivate the enhancement of usage patterns produced by clustering with qualitative and quantitative information concerning antecedents and consequents, as well as about the values of *confidence* and *support*, respectively. Consequently, the combination of clustering and association rule mining techniques may potentially result to more informative and qualitative usage patterns.

3.3 Sequential pattern mining

Sequential pattern mining was introduced in (Agrawal and Srikant, 1995), where

the problem of finding inter-transaction patterns has been defined. Under this context, a pattern is an ordered list of sets of items. According to the authors, "*given a sequence database where each sequence is a list of transactions ordered by transaction time and each transaction consists of a set of items, find all sequential patterns with a user-specified minimum support, where support is the number of data sequences that contain the pattern*". Three algorithms based on the key idea of the Apriori algorithm have been also presented in (Agrawal and Srikant, 1994). As it has already been discussed, sequential mining takes into consideration the ordering of accessed items.

Sequential mining has also been applied in the web domain. For example, we consider the following association rule:

30% of users who placed an online order in book_store/book1.html have also placed an order in book_store/book5.html within 20 days.

The problem of mining sequences of web navigational patterns refers to the identification of those web document references which are shared across time among a large number of user sequences, where a *user sequence* is a time-ordered set of visits. As an example, consider the user sequence S=<(C,D) (A,B,C) (A,B,F) (A,C,D) (E)>, where A, B, C, D, E, and F are document references. This sequence consists of 5 user visits and is called a 5-sequence. Documents C and D were accessed during the 1st visit, A, B, and C were accessed during the 2nd one, etc., while document E was the only document accessed during the 5th visit. The objective of sequential pattern mining is to enumerate the complete set of t-frequent sequences from a given database of user sequences, where t is a *minimum support threshold* (Mortazavi-asl, 2001). The resulting sequential patterns represent the web documents most frequently accessed within and across visits. (C) (A B) could be a possible pattern revealed by applying sequential pattern mining techniques in a database of user sequences. This is interpreted as following: A user of a given hypermedia system first visits document C and afterwards documents A and B. These visits need not be consecutive. Users who visited some other documents in-between also support this sequential pattern.

Most of the techniques for mining sequential patterns for WWW adopt variations of Apriori-like algorithms although they may consider different parameter settings and constraints (Han & Kamber, 2001). Some other interesting approaches have been proposed especially for mining web log data. Chen, Park, and Yu (1996) introduced the concept of "*maximal forward references*", which is defined as the sequence of documents requested by a user up to the last document before backtracking. *Maximal Forward (MF) references* algorithm aims at converting the original sequence of web server logs into a set of traversal patterns based on statistically dominant paths and association rule discovery. (Pei, et al. 2000) presented the *WAP-mine* (mining access patterns in web access sequence database) algorithm, using a conditional search strategy, based on the proposed *WAP-tree structure*. The latter is a structure facilitating web access pattern mining. On the other hand, Mortazavi-asl (2001) worked on mining sequential patterns by progressively partitioning the databases of user sequences into smaller sub-databases. They introduced the novel projection-based algorithm *PrefixSpan* for mining sequential patterns, which uses the frequent sequence lattice to partition the user sequences database.

Source	Graph structure	Input data structures	Input parameters	Algorithm	Type of Patterns
Chen et al./1996	No	Database of maximal forward references	Number *k* of references in a sequence, support threshold	Hash and pruning-based algorithms for mining association rules	Longest access sequences and/or tree patterns among users
Pei et al./2000	No	Database of web access sequences	Support threshold	Conditional search strategy (WAP-mine)	Sequential access patterns
Mortazavi-asl/2001	No	Database of web access sequences	Support threshold	PrefixSpan	Sequential access patterns
Inokuchi et al./2003	Yes	Database of web access sequences, adjacency matrix for representing graph structure data	Support threshold	Apriori-like algorithm	Subgraphs frequently included in graph structured transactions
Garofalakis et al./2002	No	Database of web access sequences	User-specified constraint *C*, support threshold	Apriori-like family of algorithms (SPIRIT family)	Frequent sequences satisfying the user-specified constraint *C*
Masseglia et al./1996	No	Database of web access sequences, hash-tree data structure	User-specified time constraints, support threshold	GENERAL algorithm	Frequent generalized sequences

Table 7. Sequential pattern mining approaches

These approaches work efficiently under some mining criteria determined by the authors. The graph structure of hypermedia systems allows the application of algorithms which have the ability to reveal patterns from graph-structured data. For this purpose, Inokuchi, Washio, and Motoda (2003) introduced a principle for deriving patterns frequently appearing in structured data. According to the authors, this approach can efficiently be applied for the analysis of users browsing behaviour.

A different approach is presented in (Garofalakis, Rastogi, and Shim, 1999), where the *SPIRIT* family of algorithms have been developed for mining frequent sequential patterns. These algorithms satisfy user-specified constraints on the mined patterns, along with the user-specified value for the support threshold. User-specified constraints are expressed with the use of a flexible constraint specification language consisting of regular expressions. As a consequence, data miners may control the mining process, while the results obtained may better satisfy their needs.

Another approach enhanced with user-specified time constraints has been proposed in (Masseglia, Poncelet, and Teisseire, 1999), where the specification of minimum and maximum time gaps among web log entries are required by the user. The processing of user access data is performed by using the *GENERAL* algorithm, a variation of the *GSP* algorithm (Srikant and Agrawal, 1996) for finding generalized sequential patterns.

Table 7 summarizes the above discussion, using the same set of parameters defined in Table 5.

4. Aspects of adaptation

As presented in the previous section, web usage mining conducts to the discovery of usage patterns, which could be *clusters of web document references*, *clusters of user visits*, *association rules among document references*, and *sequences of frequently accessed documents*. We have also discussed the special characteristics of these different approaches. In this section, we correlate the different types of usage patterns with their possible use concerning web site adaptation.

In general we can distinguish three different *aspects of adaptation* as the result of application of web usage patterns (Mobasher at al. 2001) :

- personal recommendation;
- dynamic adjustment; and
- static page/site adjustment.

In particular, *personal recommendation* concerns the adaptation of a given hypermedia system by giving advices and recommending the appropriate items to the user according to her/his user model. Recommendations can be simple propositions or even more accurate predictions of the user's next move. *Dynamic adjustment* refers to the on-line modification of the hypermedia system in a user-transparent way. Finally, *static page/site adjustment* concerns the use of patterns revealed during the web usage mining sub-process to redesign the hypermedia system.

Clusters of document references reflect patterns of common usage. This approach may be useful for all three aspects of adaptation. In the case of personal recommendation, clusters guide the online adaptation mechanism for recommending the related items, while in the case of dynamic site adjustment the clusters are used for modifying the hypermedia system structure. The active user session is associated with the appropriate cluster by computing the similarity of the documents that she/he visits during the current session with the clusters found during web mining. Thus, the adaptation is based on the value of this similarity measure. The third aspect of adaptation, i.e. the static adjustment of the hypermedia system is performed off-line. In this case usage patterns are presented to the designer who is responsible for the modification of the hypermedia system, e.g. grouping or segmentation of documents, establishment of new hyperlinks etc. In fact the off-line adjustment of the hypermedia system, presents the most common use of clustering approaches, since in this case the models luck certain important information, as for example the order of click-streams which is almost necessary for on-line recommendation.

Clusters of user visits, on the other hand, reflect patterns of common navigational behaviour and may provide an indication for users goals and motivations. Such an approach could be used for providing personal recommendations to users based on the similarity value of the currently active user session with the formed clusters of user visits, for example in e-learning and e-business environments. Similarity by itself is computed using fundamental aspects of path mining algorithms, as discussed in section 3.1. In this case, recommendations are usually advices and propositions to facilitate users browsing experience.

Similarly, an *association rule* reflects patterns of common usage –just like a cluster of document references – while in addition it provides information about the antecedent and consequents of the pattern. Association rules are often used for dynamic and static adjustment of the hypermedia system. The additional qualitative and quantitative information natively included in association rules conducts to enhanced usage patterns.

Regarding *sequential patterns*, the salient characteristic is that they reflect the usage of a time-ordered set of documents, as described in section 3. Sequential patterns have increased precision compared to the other approaches and thus could be used for personal recommendation, for dynamic adjustment as well as and static offline modification. If a rich set of patterns has been discovered a very precise and to the point navigation support may be provided during run time. However, previous navigation behaviour of users cannot usually cover all possible navigation paths, so the set of sequential patterns is often very sparse, and the technique produces low coverage of recommendations and navigation support to the users.

In summary, from the discussion of this section, we have seen that most usage patterns could be applied for different adaptation types. However, one needs to take into consideration, the distinct characteristics of each approach, in addition to other characteristics like size of the data set, coverage of usage patterns in available data, existence of good quality labelled data, structure of the hypermedia system etc. In addition, the use of other *descriptive characteristics* of the hypermedia system along with usage data could enforce the quality of the revealed patterns. These characteristics could be related with certain features of the hypermedia systems, like the topology, size, depth, degree of connectivity among the web documents, appearance, etc (Avouris, Koutri, and Daskalaki, 2003). So, user models adjust to the specific characteristics of a given hypermedia system or sub-system in order to better support adaptation. The impact of structural characteristics has also been investigated by Nakagama, and Mobasher (2003), who related several web mining approaches with the degree of connectivity and the position of the user within the hypermedia system and found that the performance of different approaches like association rules, and sequential patterns depends on the hypermedia system structure.

In summary, in Table 8 usage patterns are correlated with the most appropriate contexts of use.

Web usage mining technique	Type of usage patterns	Interpretation	Aspects of adaptation
Clustering web document references	Groups of web document references	Patterns of common usage reflecting mentally related web documents	Static page/site adjustment
Clustering user visits	Groups of user visits	Patterns of common navigational behaviour reflecting users actions and motivations	Personal recommendation
Association rule mining	Associations among web document references	Patterns of common usage reflecting related web documents, as well as the notion of antecedents and consequents	Dynamic adjustment Static page/site adjustment
Sequential pattern mining	Associations among sequences of web document references over time	Patterns of typical browsing behaviours over time	Personal recommendation

Table 8. Aspects of adaptation according to the type of usage patterns

5. Conclusions

Web usage mining has been proven a particularly suitable approach for building adaptive web systems. Web server logs constitute a rich data source allowing the experimentation with real data sets "over the user's shoulder", while guard user's privacy. A number of successful industrial applications of adaptive hypermedia systems have appeared (Brusilovsky and Maybury, 2002), mostly in the areas of education and information retrieval, while the richness and versatility of the research approaches is reflected in the survey of this chapter.

Adaptive web systems facilitate users' navigation by providing user-oriented access. During the adaptation process, user access data is the main source of information that is used for building the user model, which reflects the patterns that hypermedia systems infer for users, and describes several characteristics of users including navigational behaviour. Web usage mining concerns the application of data mining techniques specifically to the raw web access data and constitutes a challenging task for the discovery of hidden patterns.

Cluster mining, association rule mining, and sequential pattern mining are the most frequently used techniques in the context of web usage mining. This survey of several web usage mining efforts indicates the dominance of cluster mining. Traditional clustering algorithms have been modified and new ones have been invented to serve the specific needs of web usage mining. Clustering results can have the form of groups of web document references or groups of user visits. In the first case, the resulted clusters reflect patterns of common usage and reveal related web documents. In the latter case, the clusters reflect patterns of common navigational behaviour representing users' actions and motivations. Both types of clustering approaches capture a broad range of adaptation decisions, while they lack precision.

The discovery of non-sequential, as well as sequential patterns are another type of techniques for web usage mining. The discovery of association rules is actually based on traditional association rule mining algorithms, such as the Apriori algorithm. Thus, the survey of association rule mining approaches has been based on some discrete approaches. An association rule represents patterns of common usage reflecting related documents together with the notion of antecedents and consequents. Similarly to the clustering approaches, association rules are more generic and consequently less accurate.

Sequential patterns reflect the browsing behaviour of users over time, so they are capable of providing more accurate recommendations to the user, however they are more difficult to implement and require rich datasets to reflect the diversity of users behaviour.

Thus, different web usage mining techniques provide different types of usage patterns. Consequently, one needs to take into consideration the distinct characteristics of these approaches, as well as the specific characteristics of the hypermedia system and the desired adaptation level in order to select the most appropriate approach.

Current research efforts produce promising directions that refer to the *integration of web usage mining with web structure and web content mining* for augmenting the effectiveness of the resulted patterns. Towards this direction, Mobasher, et al. (2000) propose the incorporation of content models during the data processing phase. Content models refer to different ways in which web documents with similar content could be grouped together.

Semantic web mining is another active research area integrating web mining with ontologies and meta-data. Its main objective is to associate each web document with one or more ontological entities, in order to better interpret users' navigational behaviour. Berendt, Hotho, and Stumme (2002) discuss how the semantic web may improve the results of web mining by exploiting the new semantic structures on the web.

Finally, an issue of prime concern for adaptive web systems is the *evaluation of the effectiveness of adaptation*. Most studies dealing with the evaluation of adaptive web systems compare the adaptive system to the non-adaptive one or apply traditional machine learning evaluation techniques, which do not involve users. However there are many limitations to these approaches. For instance, comparing of adaptive and non-adaptive systems is not fair as the non-adaptive instance cannot be optimal, if adaptation is properly designed into the system (Höök, 2000). Several traditional empirical usability evaluation methods have also been used in the context of adaptive web systems. Interviews, questionnaires, think aloud protocols, feature checklists constitute some frequently used methods for the investigation of usability. Another interesting direction is the evaluation of the usage of the web system by applying web usage mining techniques (Koutri, and Daskalaki, 2003). Processing user access data may help web miners understand how the web site is perceived by the users. This knowledge is not trivial to obtain, because the web designer has a different mental model of the system. Consequently, the discovery of usage patterns and their interpretation provides useful results concerning users' navigational patterns through the given hypermedia system. However, more research and practice is needed in this area in order to obtain more solid methods and techniques.

6. Acknowledgments

This study has been performed in the frame of the project "*Development of probabilistic models of web use*", funded by the UoP Research Committee K. Karatheodoris basic research program.

7. References

Agarwal, R., Aggarwal, C., and Prasad, V. (1999). A Tree Projection Algorithm for Generation of Frequent Itemsets. Proceedings of High Performance Data Mining Workshop.

Agrawal, R., and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th Very Large Database (VLDB) International Conference, 487-499.

Agrawal, R., and Srikant, R. (1995). Mining Sequential Patterns. Eleventh International Conference on Data Engineering, IEEE Computer Society Press, 3-14.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Peter Buneman and Sushil Jajodia, 207-216.

Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T. (1995). WebWatcher: A Learning Apprentice for the World Wide Web. AAAI Spring Symposium on Information Gathering, 6-12.

Avouris N., Koutri M., and Daskalaki S. (2003). Web site adaptation: a model-based approach. Proceedings of HCII 2003 Human-Computer Interaction 2, Lawrence Erlbaum Assoc., Mahwah, NJ, 350-354.

Banerjee, A., and Ghosh, J. (2001). Cliskstream Clustering using Weighted Longest Common Subsequences. Proc. of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 33-40.

Berendt, B., Hotho, A., and Stumme, G. (2002). Towards Semantic Web Mining. International Semantic Web Conference (ISWC02).

Borges, J., and Levene, M. (1998). Mining Association Rules in Hypertext Databases. Knowledge Discovery and Data Mining, 149-153.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. User Modeling and User-Adapted Interaction 6, 2-3, 87-129.

Brusilovsky, P., and Maybury, M. T. (2002). From adaptive hypermedia to the adaptive web. Communications of the ACM 45, 5, 30-33.

Chen, M. S., Park, J. S., and Yu, P.S. (1996). Efficient Data Mining for Path Traversal Patterns. IEEE Trans. on Knowledge and Data Engineering 10, 2, 209-221.

Cooley, R., Mobasher, B., and Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems 1, 1, 5-32.

Foss, A., Wang, W., and Zaiane, O. R.(2001). A non-parametric approach to web log analysis. Proceedings of Workshop on Web Mining in First International SIAM Conference on Data Mining, Chicago, 41-50.

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

Garofalakis, M., N., Rastogi, R., and Shim, K. (1999). SPIRIT: Sequential Pattern Mining with Regular Expression Constraints. The VLDB Journal, 223-234.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. (1997). Data Cube: A Relational Aggregation Operator Generalizing Group-By. Data Mining and Knowledge Discovery 1, 1, 29-53.

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

Hanson, R., Stutz, J., and Cheeseman, P. (1991). Bayesian classification theory. Technical Report FIA-90-12-7-01, AI Branch, NASA Ames Research Center, CA.

Hipp, J., Guntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining-A General Survey and Comparison. SIGKDD Explorations 2, 1, 58-64.

Höök, K. (1998). Evaluating the Utility and Usability of an Adaptive Hypermedia System. Knowledge Based Systems, 10, 5.

Inokuchi, A., Washio, T., and Motoda, H. (2003). Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. Machine Learning, 50, 3, 321-354.

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

Kamba, T., and Sakagami, H. (1997). Learning Personal Preferences on online Newspaper articles from user behaviours. Proc. 6th Int. World Wide Web Conference.

Karypis, G., and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. on Scientific Computing, 20, 1, 359-392.

Kohonen, T. (1997). Self-organizing maps. Berlin: Springer Verlag, 2nd ed..

Koutri, M., and Daskalaki, S. (2003). Improving web site usability through a clustering approach. Proceedings of HCII 2003 Human-Computer Interaction 1, Lawrence Erlbaum Assoc., Mahwah, NJ, 788-792.

Lieberman, H. (1995). Letizia: An agent that assists web browsing. Proc. 14th Int. Joint Conference on Artificial Intelligence (IJCAI95), Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 924-929.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proc. 5th Berkley Symposium on Mathematical Statistics and Probability, I: Statistics, 281-297.

Masseglia, F., Poncelet, P., and Teisseire, M. (1999). Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure. ACM SigWebLetters 8, 3, 13-19.

Minio, M., and Tasso, C. (1996). User Modeling for Information Filtering on INTERNET Services: Exploiting an Extended Version of the UMT Shell. Proceedings 5th Int. Conference on User Modeling, Kailia-Kuna, Hawaii.

Mobasher, B., Berendt, B., and Spiliopoulou, M. (2001). KDD for Personalization. PKDD 2001 Tutorial.

Mobasher, B., Cooley, R., and Srivastava, J. (1999). Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. IEEE Knowledge and Data Engineering Workshop (KDEX'99).

Mobasher, B., Dai, H., Luo, T., Sung, Y., and Zhu, J. (2000). Integrating Web Usage and Web Content Mining for More Effective Personalization. Proc. Int. Conference on E-Commerce and Web Technologies (ECWeb2000).

Mobasher, B., Jain, N., Han, E., and Srivastava, J. (1996). Web mining: Pattern discovery from world wide web transactions. Technical Report TR-96050, Department of Computer Science, University of Minnesota, Minneapolis.

Mortazavi-asl, B. (2001). Discovering and Mining User Web-Page Traversal Patterns. Master Thesis, School of Computer Science, Simon Fraser University.

Nakagawa, M., and Mobasher, B. (2003). A Hybrid Web Personalization Model Based on Site Connectivity. Proc. 5th WEBKDD workshop: Webmining as a Promise to Effective and Intelligent Web Applications (WEBKDD'2003), 59-70.

Nasraoui, O., Krishnapuram, R., Frigui, H., and Joshi, A. (2000). Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering. Int. J. on Artificial Intelligence Tools 9, 4, 509-526.

Paliouras G, Papatheodorou C, Karkaletsis V, and Spyropoulos, C. D. (2000). Clustering the Users of Large Web Sites Into Communities. Proc. 17th Int. Conf. on Machine Learning (ICML), Morgan Kaufmann, San Francisco, CA, 719-726.

Pazzani, M. J., and Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning 27, 3, 313-331.

Pei, J., Han, J., Mortazavi-asl, B., and Zhu, H. (2000). Mining Access Patterns Efficiently from Web Logs. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00), 396-407.

Perkowitz M, and Etzioni, O. (1997). Adaptive Web Sites: An AI challenge. Proc. 15th Int. Joint Conference on Artificial Intelligence (IJCAI), 16-23.

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

Pierrakos, D., Paliouras, G., Papatheodorou, C., and Spyropoulos, C.D. (2003). Web Usage Mining as a Tool for Personalization: A Survey. User Modeling and User-Adapted Interaction 13, 311-372.

Shahabi, C., Zarkesh, A.M., Adibi, J., and Shah, V. (1997). Knowledge Discovery from User Web-Page Navigation. Proceedings of Workshop on Research Issues in Data Engineering (RIDE), Birmingham, England, 20-29.

Srivastava J, Cooley R., Deshpande M., Tan P., (2000). Web usage mining: Discover and applications of usage patterns from web data, SIGKDD Explotations, 1(2), 12-23.

Srikant, R., and Agrawal, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. Proceedings of the 5th Int. Conference Extending Database Technology 1057, 3-17.

Wang, W., and Zaiane, O.R. (2002). Clustering Web Sessions by Sequence Alignment. Proceedings of 13th Int. Workshop on Database and Expert Systems Applications (DEXA'02), Aix-en-Provence.

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

Webb G., Pazzani M., Billsus D., (2001). Machine Learning for User Modeling, User Modeling and User-adapted Interaction, 11, 19-29.

Wong, C., Shiu, S., and Pal, S. (2001). Mining Fuzzy Association Rules for web access case adaptation.

Yan, T., Jacobsen, M., Molina, H., and Dayal, U. (1996). From User Access Patterns to Dynamic Hypertext Linking. Proc. 5th Int. World Wide Web Conference, Paris, France.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996), BIRCH: An Efficient Data Clustering Method for Very Large Databases. Proc. of ACM SIGMOD Conference on Management of Data, 103-114.