CHAPTER 4

On using data mining for browsing log analysis in learning environments

F. Wang
Department of Computer Science and Information Engineering, Ming Chuan University, Taiwan, R.O.C.

Abstract

Recently, the rapid progress of Internet technology has triggered the widespread development of web-based learning environments in the educational world. As compared with conventional CAI systems, web-based learning environments are able to accumulate a huge amount of learning data. As a result, there is an urgent need for analyzing methods of discovering useful knowledge from the huge log database for improving instructional/learning performance. In this chapter, I will present some models and methods of analyzing the browsing log data to construct a browsing behavioral model that is helpful in supporting e-learning applications. For example, teachers can investigate the model to identify some interesting or unexpected learning patterns in student's browsing behavior, which might therefore provide knowledge for teachers to reorganize their content structure in a more effective manner. Alternatively, another model can be used as a reference model by which personalized content recommendation could be made. To serve these purposes, a set of tools based on data mining techniques such as clustering, and association mining, combined with collaborative filtering techniques, are developed. The effectiveness of these methods is investigated on a real database collected from web-based courses. Through the case studies, some revelations are presented and some future research directions are discussed.

1 Introduction

Web-based learning environments have been the main trend for technology-enhanced education in the last decade. In this new era of web-based education

WIT Transactions on State of the Art in Science and Engineering, Vol 2, © 2006 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/1-84564-152-3/04

technology, the Internet and the World Wide Web have been exploited as a vast repository of information, playing the role of providers of educational resources. In web-based learning environments, teachers could conduct many kinds of instructional/learning activities such as online material browsing, exercise practicing, group discussion, online testing and so on. However, one of the prevailing issues in such a learning environment is that it is not easy to monitor students' learning behavior. Nevertheless, as compared with conventional CAI systems, web-based learning environments are able to keep track of most learning behavior of the students, and hence are able to provide a huge amount of learning profiles. As a result, there is an urgent need of analyzing methods to discover useful information for improving instruction/learning performance from the huge log database. These learning profiles provide teachers a valuable data source to observe and analyze students' learning processes and performance [1, 2].

Data mining in e-learning has been receiving more and more attention from researchers in various learning aspects. For example, Tang *et al.* [3] proposed a technique to construct personalized courseware based on data mining. Abramowicz *et al.* [4] used data mining techniques to support the creation of topic map so that distributed content resources could be shared and reused more efficiently. Zaiane exploited web access logs and advanced data mining techniques to extract useful patterns that can help instructors evaluate and interpret online course activities to assess the learning process and measure web course structure effectiveness [5, 6]. Personalized e-learning through delivering personalized content has been one of the main focus of research and has receiving many interesting results [7–10].

While there are several kinds of online learning activities, this chapter focuses on the online material browsing behavior. Traditionally, online materials are divided into units of topics that are organized and structured by some semantic relations among themselves [11]. Analyzing browsing behavior of students in a web-based learning environment might reveal some insights into the true structure required of the online material for being helpful to student learning. Such knowledge about the dynamic browsing structure would be an important reference base in designing effective online educational materials. It may also be helpful to the design of adaptive navigation guiders and/or personalized recommenders. Therefore, this chapter proposes a research framework in which browsing log can be dealt with in such a way that the aforementioned e-learning applications can be improved. Specifically, a set of browsing models to describe useful browsing patterns are proposed, and analysis tools based on web mining technique [12] to discover those patterns from the historical browsing database will be presented.

Finally, applications of the analysis methods are conducted on a real database collected from three web-based courses at Ming Chuan University, Taiwan. Three classes of the Expert System course had been conducted for a semester at Ming Chuan University, Taiwan; two of the courses were open to daytime students and the third to on-job students. Students were grouped and they were required to do a term project of building an expert system. During the semester, students had to work collaboratively in a web-based virtual classroom, in which their interactions and activities such as online material browsing were recorded in a back-end database

for both evaluation and investigation purposes. The analysis results illustrate the potential capability of the methods to reveal useful browsing knowledge as a basis for investigation and comparison of student's learning behavior.

2 Data mining

Data mining, which is also referred to as knowledge discovery in database, is a process of non-trivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in database [13]. The data mining algorithms can be divided into three major categories based on the nature of their information extraction: predictive modeling (also called classification or supervised learning), clustering (also called segmentation or unsupervised learning), and frequent pattern extraction [14]. In the following, we briefly review some of the mining methods and applications that are relevant to our research.

2.1 Association mining

Association mining is one of the most well studied mining methods in data mining [13–16]. It serves as a useful tool for discovering correlations among items in a large database. It explores the probability that when certain items are present, which other items are also present in the same affairs. An association rule is a condition of the form $X \Rightarrow Y$ where X and Y are two sets of items. An interpretation of the association rule in a business trade situation is when a customer buys items in X, the customer will also buy items in Y.

There are two important threshold values used in mining association rules: *support* and *confidence*. Support indicates the frequencies of the occurring patterns in the rule. In the minimum support approach, association rules are generated by discovering *large itemsets*. A set of items X is called a large itemset if the support rate of X, with respect to a transaction database, meets the minimum support requirement. Confidence denotes the strength of the implication of the association rule. If the confidence is higher, the rule is more reliable.

2.2 Clustering

Clustering is a useful technique for discovering interesting data distributions and patterns in the underlying data. It is a process of grouping physical or abstract objects into classes of similar objects. Clustering analysis helps construct meaningful partitioning of a large set of objects based on a 'divide and conquer' methodology which decomposes a large scale system into smaller components to simplify design and implementation [13]. The principle of clustering is maximizing the similarity inside an object group and minimizing the similarity between the object groups.

The most well-known and commonly used partitioning methods are k-means, k-medoids, and their variations [16]. In the k-means algorithm, cluster similarity is measured in regard to the mean value of the objects in a cluster, which can

be viewed as the cluster's center of gravity. The k-means method, however, can be applied only when the mean of a cluster is defined. This may not be the case in some applications, such as when data with categorical attributes are involved. Besides, it is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data. On the other hand, instead of taking the mean value of the objects in a cluster as a reference point, the k-medoids method use the medoid, which is the most centrally located object in a cluster. Therefore, the k-medoids method takes advantage over the k-means in the aspects of versatileness and outlier insensitivity. However, the necessity of both methods for users to specify k, the number of clusters, in advance can be seen as a common disadvantage.

2.3 Web usage mining

In the World Wide Web context, web sites are generating a great amount of web usage data that contain useful information about users' behavior. The term 'web usage mining' was introduced by Cooley *et al.* [12], in 1997, in which they define web usage mining as the 'automatic discovery of user access patterns from web servers'. Web usage mining has gained much attention in the literature as a potential approach to fulfilling the requirement of web personalization [12, 17–21]. The discovered knowledge indicating users' navigational behavior is useful for the system to personalize the web site according to each user's behavior and profile. The data mining methods that are employed including association rule mining, sequential pattern discovery, clustering and classification. In this chapter we focus on the association mining method, which is a widely used data analysis method in web usage mining [7, 8, 17, 22].

3 Recommendation systems

In a large-scale distributed network environment like the Internet, the popularization of computers and the Internet have resulted in an explosion in the amount of digital information. As a result, it becomes more important and difficult to retrieve proper information adapted to user preferences. Therefore, personalized recommendation systems are in need to provide proper recommendations based on users' requirements and preferences [21, 23] In general, there are two types of recommendation systems, the content-based filtering systems and the collaborative filtering systems [20, 24].

3.1 Content-based filtering systems

Content-based filtering techniques are based on content analysis of target items. For examples, the technique of term frequency analysis for text document and its relation to the user's preferences is a well-known content analysis method. In content-based filtering systems, recommendations are provided for a user based

solely on a profile built up by analyzing the content of items that the user has rated in the past and/or user's personal information and preferences. The user's profile can be constructed by analyzing the responses to a questionnaire, item ratings, or the user's navigation information to infer the user's preferences and/or interests. However, a pure content-based filtering system has several shortcomings and critical issues remained to be solved, including that only a very shallow analysis of specific kinds of content (text documents, etc.) are available and that users can receive only recommendations similar to their earlier experiences and the sparseness problem of item rating information [22, 25].

3.2 Collaborative filtering systems

In collaborative filtering, items are recommended to a particular user when other similar users also prefer them. The definition of 'similarity' between users depends on applications. For example, similarity may be defined as users having similar ratings of items or users having similar navigation behavior. This kind of recommendation system is the first that uses the artificial intelligence technique to do the personalized job [23]. A collaborative filtering system collects all information about users' activities on the web site and calculates the similarity among the users. If some users have similar behavior, they will be categorized to the same user group. When a user logins into the web site again, the system will first compute the group most similar to the user using methods like the k-nearest neighborhood, and then recommend items that the members of the group prefer to the user. A pure collaborative filtering system also has several shortcomings and critical issues, including that the coverage of item ratings could be very sparse, hence yielding poor recommendation efficiency; and that it is difficult to provide services for users who have unusual tastes, and the user clustering and classification problems for users with changing and/or evolving preferences [26]. Table 1 shows a brief comparison between the two filtering methods.

3.3 Recommendation systems based on association rules mining technologies

As data mining techniques become more and more mature, researchers have explored their applications in recommendation systems in the last decade, trying to improve the efficiency and the effectiveness of the recommendation systems. Among those efforts, Fu *et al.* [19] try to integrate the collaborative filtering method and association mining technology to develop a recommendation system called SurfLen that recommends web pages on the web site. Their research reorganized the web pages collected from the 'Yahoo!' search engine, and experimented on the influence of the noise upon the recommendation effectiveness [19]. Besides, Lee *et al.* [22] integrate the collaborative filtering method and association mining technology to develop a recommendation system to recommend movies for the audiences on the MovieLens web site (http://www.movielens.umn.edu).

Table 1: Comparison between content-based filtering and collaborative filtering systems [9].

	Content-based filtering	Collaborative filtering
Advantage	1. A user can receive proper recommendations without help from other users.	1. A user may have a chance to receive items that s/he never contacted before, but may be of his/her potential interest.
	2. It is more feasible to tackle the problems of multiple user interests and interest transference by monitoring the change and evolving of user profiles.	2. Facilitate the sharing of knowledge and/or experiences among users having similar interests.
Limitation	1. Some types of items (e.g. multimedia) are not easy to analyze.	1. It is hard to provide recommendations for users who have unusual preferences.
	2. A user can just receive items that are similar to his/her past experiences.	2. It is hard to cluster and classify users with changing and/or evolving preferences.

4 The research framework

Due to the rapid growth of e-learning applications on the Internet, the complexity of the tasks such as content structure design, LMS server design, and content navigation design has increased along with this growth. To handle these complex tasks, we need knowledge about user behavior characteristics. This section presents a research framework that integrates data mining techniques to extract knowledge for specific e-learning applications from user's historical activity, in particular for intelligent personalized services.

As shown in Fig. 1, the research framework consists of six research tasks that have to be dealt with properly. The first is learning activity design, which deals with the problem of designing learning tasks that a researcher is interested. For example, material browsing is a common learning task that in the literature researchers have been most interested. The design of learning activity depends on the purposes of the specific e-learning applications. For example, to facilitate the design of proper material structure, dynamic document browsing model could be constructed from the browsing history data such that the browsing patterns can be reflected in an improved content structure. Another example of e-learning application is the matching of a student with some other well-performing group of users that share

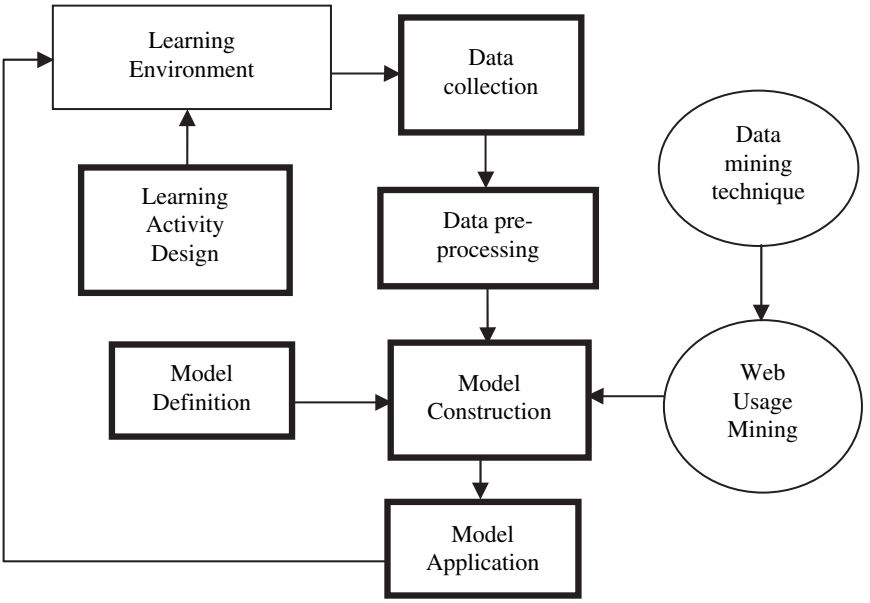

Figure 1: A research framework for data mining and e-learning.

similar activity characteristics such that their behavior can be referenced to give proper suggestions to the student. User browsing model could be constructed from logged user browsing data that could reflect the real browsing behavior of well-performing users. In summary, what learning activities are provided to students need to be decided first before we can go on to the next step.

Data tracking is required to facilitate the data collection for the interested learning tasks. A designer of data tracking needs to select proper tracking attributes such as the user id, the date and time period of the learning task a user performed, and so on. For example, in the aforementioned example of matching a student with some other well-performing group of users that share similar activity characteristics, the categories of activities performed by a student and their performance results are both required to be tracked for further analysis.

Data pre-processing is one of the important research tasks in this field. Data has to be cleaned and transformed properly before it can be analyzed. In some situations where user sessions are not easy to identify, several heuristic methods have been explored to decide user sessions from the logged access history (often logged in an http server). In some other cases where registered user login operation is required, user session determination is not a problem. Nevertheless, determination of meaningful user sessions is still a challenging problem. Besides, since it often happens that students may navigate documents back and forth, we need a way to handle a tree-structured browsing behavior such that the user's meaningful intentions in a browsing session can be identified. Some researchers have proposed heuristic

methods to divide a user session into a set of shorter meaningful sub-sessions. These sub-sessions are the real sessions fed into the next step of model construction using data mining techniques.

The next step is the model definition and construction. Researchers have to define a behavior model that reflects the real topics or issues they are interested. For example, a browsing model that depicts the real browsing behavior can be defined in terms of the association and sequential browsing patterns that occur often in the history. These models can be used by teachers to identify some interesting or unexpected learning patterns in student's browsing structure, and therefore might provide knowledge for teachers to reorganize their content structure in a more effective manner. After the model definition, efficient and effective methods to construct the model are to be derived next. In this chapter, I will focus on web usage mining and collaboration filtering techniques that can be of help in this phase. Finally comes the model application phase to investigate the fitness and effectiveness of the devised model. Some evaluation metrics have to be defined. In the following, I will present two cases studies of our previous work based on this research framework. One is for content structure model construction [9], and the other is for navigation guidance by personalized recommendation [27].

5 Construction of browsing content structure

Traditionally, web materials are divided into units of topics that are organized and structured by some semantic relations among themselves [11]. However, the organization of online content does not necessarily meet the individual needs of students. On the other hand, as the content volume increases with time, maintenance of the content structure may become an uneasy load for designers. Analyzing browsing behavior of students might reveal some insights into the real content structure for being helpful to student learning. Such knowledge about the dynamic browsing structure would be an important reference base for designing more effective online materials. This section presents a browsing model to describe useful browsing patterns, and develops analysis tools based on data mining technique to discover those patterns from the historical browsing log.

5.1 Data pre-processing

The data analysis process consists of the following five stages: (1) data filtering, (2) data transformation, (3) frequency analysis, (4) co-reference mining of document clusters, and (5) sequence mining, as shown in Fig. 2. Since this study focuses on the browsing activities, all other unrelated data are filtered out, including those browsing records with short stay-time, e.g. reference pass-by pages. To validate long-stay-time records, a client program could be deployed to monitor users' interactions with the computer. Besides, the raw data has to be reconfigured for each student's browsing session. First, all the browsing records are sorted with the student id as a major key and start time as a minor key in an ascendant manner.

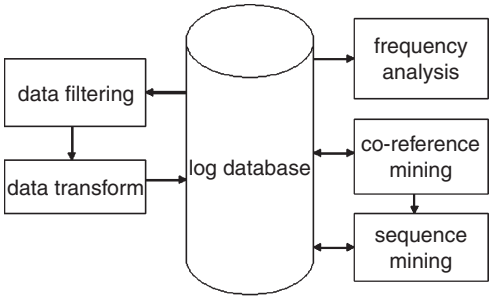

Figure 2: The data analysis process of content-structuring browsing model.

Then, browsing records picked up between two successive 'login' records are grouped into a browsing-session record. As it often happens that students navigate documents back and forth, we need a way to handle such a tree-structured browsing behavior [28]. This study adopts the pre-order scan approach [25] for converting student's tree-structured navigation paths so that a maximal number of ordered browsing sequences could be obtained.

Frequency analysis is helpful for instructors to get an overview of the usage of various categories of materials. Two kinds of frequency analysis may be performed: (1) the hit rate analysis of the material, and (2) the summary hit rate analysis of the material categories. Co-reference clustering and sequence mining are then performed to construct the content structuring model.

5.2 Model definition and construction

The patterns we are interested here include the co-referenced document clusters and sequences between the clusters. A document cluster indicates a set of documents that students often study together (i.e. the co-referenced knowledge units), while the sequence rules between the document clusters reflect students' knowledge construction sequence in a specific target domain, which also reflects possible prerequisite relations between the documents. Associative mining technique can be applied to find document clusters. Content documents can be bound together with different strength. By adopting different support rates, the documents form a hierarchical cluster structure so that teachers can investigate document clusters in different grain size (i.e. association strength). Then a sequence mining (of length 2) can be applied to find the binary sequence patterns between document clusters.

The following definitions are adapted from [27] to describe the content structure model formally. First, a learning session is the duration when a student logs into the system until he/she leaves the system. Since we are only interested in browsing activity, each session retains only browsing activities by removing other learning activities present in it.

Definition 1 (Co-reference relation): For any documents P_1 and P_2, we say P_1 is co-referenced with P_2 if and only if P_1 and P_2 are 'frequently' browsed in a learning session without 'observable' ordering between P_1 and P_2.

Definition 2 (Sequence relation): For any documents P_1 and P_2, we say P_1 precedes P_2 if and only if P_1 and P_2 are 'frequently' browsed together and there is 'observable' ordering between P_1 and P_2.

Definition 3: Given a pair of documents (P_1, P_2), define the 'sequence strength' of $P_1 \rightarrow P_2$ as $\|P_1 - P_2\| = \max\{\text{SEQ}(P_1, P_2) - \text{SEQ}(P_2, P_1)/N(P_1, P_2), 0\}$, where $N(P_1, P_2)$ is the number of session records containing both P_1 and P_2, and SEQ (P_i, P_j) denotes the number of session records where P_i precedes P_j.

Definition 4: Given a document cluster C of size n, the intra-cluster sequence strength of C is $\|C\| = \max_{i=1,\ldots,n,\, j=i+1,\ldots,n} \|P_i - P_j\|$, where $P_i, P_j \in C$.

Definition 5: Given a set of cluster C of n documents $\{P_1, P_2, \ldots, P_n\}$, the support rate of cluster C, $\sup(C) = N(P_1, P_2, \ldots, P_n)/T$, where $N(P_1, P_2, \ldots, P_n)$ is the count of session records containing $P_1, P_2, \ldots, P_n$, and T is the total number of session records.

Definition 6: A document cluster C is a valid co-reference cluster if and only if $\sup(C) \geqq \beta$ and $\| C \| \leqq \delta$, where β is a given support rate threshold, and δ is an intra-cluster sequence strength threshold.

Definition 7: Given two document clusters C_1 and C_2, define the *support* of the sequence rule $C_1 \rightarrow C_2$ as $\sup(C_1 \rightarrow C_2) = N(C_1 \rightarrow C_2)/T$, where $N(C_1 \rightarrow C_2)$ is the number of sessions containing the pattern of $C_1 \rightarrow C_2$, and T is the total number of sessions.

Definition 8: Given two document clusters C_1 and C_2, define the *confidence* of the sequence rule $C_1 \rightarrow C_2$ as $|\sup(C_1 \rightarrow C_2)|/|\sup(C_1)|$.

Consider the browsing log shown in Table 2. Set the minimum support rate and maximum intra-cluster distance to 0.5 and 0.1, respectively. The co-reference mining proceeds as follows. Initially, all 1-clusters (i.e. the large 1-itemsets) with sufficiently large supports are calculated, resulting in four clusters $\{P_1\}$, $\{P_2\}$, $\{P_3\}$ and $\{P_4\}$. Next, the mining proceeds to discover clusters of size 2. All candidates of 2-clusters are generated by combining pair-wisely the 1-clusters, and then each checked the validity by computing its support and intra-cluster sequence strength. The resulting clusters are listed in Table 3. However, by checking both the minimum support and the maximum sequence strength constraints, only the two clusters $\{P_1P_2\}$ and $\{P_3P_4\}$ are valid. Continuing this process, a new candidate cluster $\{P_1P_2P_3P_4\}$ is produced with support 0.66 and intra-cluster sequence strength 1.

Table 2: An example log of browsing sessions.

Session #	Browsing path
1	$P_1P_2P_3P_4$
2	P_4P_2
3	$P_2P_1P_4P_3$
4	P_3P_1
5	$P_2P_1P_4P_3$
6	$P_1P_2P_3P_4$

Table 3: Support rates and intra-cluster sequence strength of 2-clusters for the log of Table 2.

Cluster	Support rate	Cluster distance
P_1P_2	0.66	0
P_1P_4	0.66	1
P_2P_3	0.66	1
P_3P_4	0.66	0
P_1P_3	0.83	0.2
P_2P_4	0.83	0.6

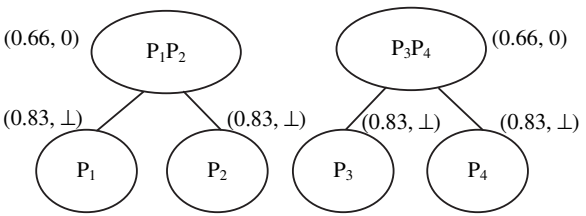

Figure 3: Hierarchical clustering of co-reference documents with (support rate, intra-cluster sequence strength).

However, it is not valid due to its intra-cluster sequence strength. As a result, the final co-referenced clusters are $\{P_1P_2\}$ and $\{P_3P_4\}$.

The co-referenced clusters $\{P_1P_2\}$ and $\{P_3P_4\}$ form a hierarchical clustering of documents with different levels of support rates, as shown in Fig. 3. A hierarchical sequence rule structure can be derived by composing the clusters in $\{P_1P_2\}$ and $\{P_3P_4\}$. An efficient lattice-product algorithm was devised that outputs all feasible sequence rules in the form of a lattice hierarchical structure [27], as shown in Fig. 4. By 'feasible' rules we mean those rules with sufficient support and confidence (say, with a minimum support and confidence constraint, say 0.5 and 0.8, respectively). Note that a link from a lower node X to its upper parent node Y indicates

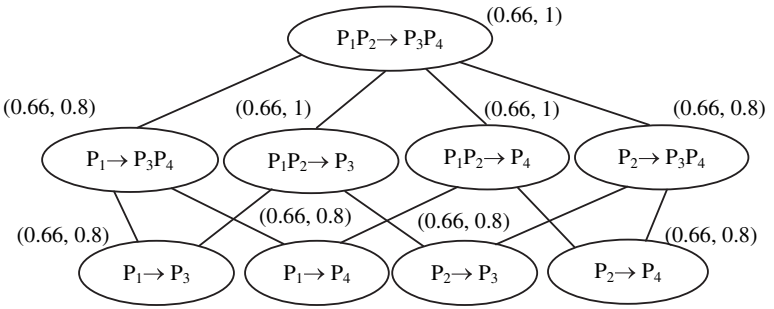

Figure 4: Hierarchical sequence rules generated from clusters in ($\{P_1P_2\}$, $\{P_3P_4\}$) with (support rate, confidence).

that the rule corresponding to node X is a generalizer of the one corresponding to node Y. For example, the rule $\{P_1\} \rightarrow \{P_3P_4\}$ is a generalized rule of the rule $\{P_1P_2\} \rightarrow \{P_3P_4\}$.

5.3 Model application

The aforementioned analysis process is performed in three web-based classes of the Expert System course at Ming Chuan University, Taiwan. A total of 172711 browsing records are stored in the log database. The minimum threshold of browsing time is 10 s and the minimum support is 0.03. A total of 22846 sessions are left after the cleaning process and 960 browsing sessions are attained after the transforming process. The total number of material documents is 96.

Through frequency analysis it was found that students in all three classes spent more efforts in browsing material of the Design category, which might have something to do with the term projects that are required in all classes. As to the Theory and Demo Systems categories, both daytime classes prefer studying the Demo System category than the Theory one. On the contrary, the on-job class prefers studying the Theory category than the Demo System. Finally, all three classes reveal similar browsing patterns of material in the Design category; i.e. the most on 'Language', then the 'Operations', and the least on 'Samples'. This implies that the original intention of providing samples to help students lean the design task more efficiently is not effectively fulfilled, and it deserves more investigation to explore the reason why.

To improve the reliability of the mining results by increasing the amount of historical data, sessions of the three classes are collected for the co-reference and sequence rule mining tasks. The thresholds of the support rate, intra-cluster sequence strength and cluster support rate are set to 0.03, 0.2 and 0.1, respectively, and the sequence strength threshold is set to 0.2. The results show that 23 clusters are found with the largest cluster being of size 3. Among these clusters, it is found that 'Theory' and 'Demo Systems' categories are often browsed together, and 'Demo System' and

'Language' categories are also often browsed together. This implies that the goal of encouraging cross-references of the materials in 'Demo System' and 'Language' was effectively achieved.

On the other hand, most of the sequence rules meet the instructor's expectations. Nevertheless, some interesting sequence patterns are also found. For example, the sequence rule ([Backward-chaining ES], [Inference Engines]) → ([CLIPS Language]) with confidence 0.3 reveals that some students misunderstood the 'CLIPS language' (a tool for designing forward-chaining expert systems) as a candidate design language for backward-chaining expert systems. This knowledge of phenomena could be used by instructors to clarify the usage of expert tools in future courses.

5.4 Summary statements

In this case study, a content structure modeling process is presented, and a tool for analyzing the historical browsing data is presented. There is more work worth further pursuit. For example, there is still a lack of more effective mining tools for analyzing other kinds of learning information, such as the behavior of 'thinking order' in a web-based online discussion context, and also those tools for helping teachers to explore the relationships between the various learning patterns and the learning outcomes [29]. Teachers could then answer the questions such as 'what are the behavioral characteristics of students tending to good learning outcomes?'

6 Personalized recommendation based on association mining

Personalized recommendation by predicting user-browsing behavior using association mining technology has gained much attention in web personalization research area [11, 21, 22, 30, 31]. In particular, it can be a potential way to create personalized e-learning applications [2, 3, 9, 20, 31, 32]. However, the association patterns did not perform well in prediction of future browsing patterns due to the low matching rate of the association rules against users' browsing behavior. According to the evaluation results of [17], the accuracy and coverage rate of the association mining technique is usually quite low. Also note that their results showed that though the sequence mining method produced higher accuracy than association mining did, it produced much lower coverage. Besides, Wang and Thao [8] applied the association mining on the whole navigation sessions to establish a knowledge model to predict users' next request in an e-learning web site. However, their results also revealed similar evidences to this fact. This drawback of applying pure association mining shows the potential limitation of the prediction knowledge built through conventional association mining technique.

Instead of performing association mining on users' navigation sessions as a whole, which might eliminate the visibility of important access patterns due to the large population, users are clustered elaborately by sampling the navigation

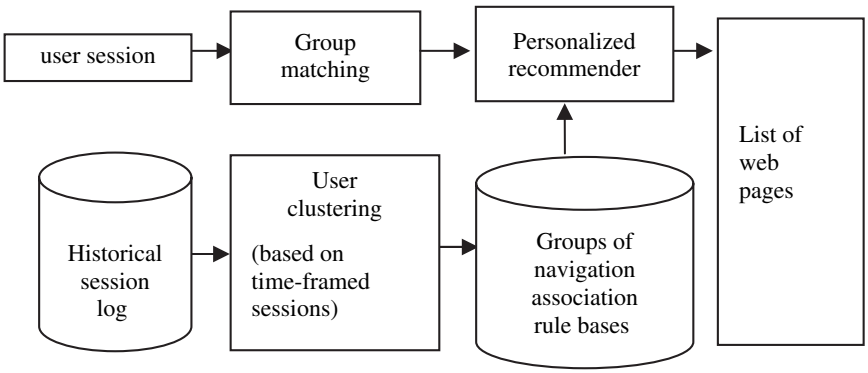

Figure 5: The personalized recommendation mechanism based on association mining and time-framed user session clustering.

sessions in a specific time frame. Wang and Shao [9] found that by clustering users properly according to their browsing behavior within specific time intervals, the recommendation effectiveness could be improved significantly. The case study presented as follows is adapted from the work of [9].

The framework of the personalized recommendation of [9] based on association mining and time-framed user session clustering is shown in Fig. 5. Users are clustered based on so-called time-framed navigation sessions, and then access patterns are discovered for each user by the association mining technique. To produce personalized recommendations for a user, the group most similar to the user's navigation sessions is first selected, and then the recommender applies the prediction rules in the corresponding rule base to generate the item recommendation list that sorts the items in terms of relevance.

6.1 Model definition and construction

Historical navigation sessions for each user are divided into frames of sessions based on a specific time interval. Selection of a good time interval is an elaborative decision that depends on the characteristics of the applications. For example, in this case study, candidate time intervals may be a 'week' or a 'semester', which coincides with the teaching/learning schedule of the testing courses.

The different impacts of the chosen time frame size are depicted as below. A long time interval, such as a 'semester', provides a macro view of a user's navigation behavior embedded with richer long-term access information, but it may be hard to generalize the navigation rules in such a macro behavior view. On the other hand, a shorter time interval provides a micro view on a user's navigation behavior with more focused access information, but it may lose long-term access information such that it is harder to perform a trend analysis.

6.1.1 User browsing similarity in time-framed navigation sessions

As mentioned above, users' navigation sessions are grouped into session frames according to a pre-specified time interval. Consider two time-framed navigation sessions from different users U_i and U_j, as shown below, respectively,

$$U_i : TF_u(U_i) = \{S_{i1},\ S_{i2},\ \ldots,\ S_{in}\}, \text{the } u\text{th time - framed sessions},$$
$$U_j : TF_v(U_j) = \{S_{j1},\ S_{j2},\ \ldots,\ S_{jm}\}, \text{the } v\text{th time - framed sessions},$$

where session S_k is a collection of web pages that the users have visited during a session at specific time interval. Then two users are said to be similar to each other in two time intervals if the two users have similar navigation behavior during the two time intervals (may be the same time interval). Specifically, define the similarity of two session records, S_{is} and S_{jt}, as follows:

$$\text{Sim}(S_{is}, S_{jt}) = \frac{\mid S_{is} \cap S_{jt} \mid}{\mid S_{is} \cup S_{jt} \mid}, \quad 1 \leqq s \leqq n, \quad 1 \leqq t \leqq m. \tag{1}$$

Next, define the similarity of two time-framed sessions $\text{TF}(U_i)$ and $\text{TF}(U_j)$ as:

$$\text{Sim}(TF_u(U_j), TF_v(U_j)) = \min(\overline{S}_{ij},\ \overline{S}_{jt}), \tag{2}$$

where $\overline{S_{ij}} = \text{Avg}\, s = 1, \ldots, n\ (\max t = 1, \ldots, m\ \{\text{Sim}(S_{is}, S_{jt})\})$ and $\overline{S_{ji}} = \text{Avg}\, t = 1, \ldots, m\ (\max s = 1, \ldots, n\ \{\text{Sim}(S_{is}, S_{jt})\})$. Actually, for the two time intervals u and v, $\overline{S_{ij}}$ indicates the average degree to which user i is similar to user j, while $\overline{S_{ji}}$ is the average degree to which user j is similar to user i, and $\text{Sim}(\text{TF}_u(U_i), \text{TF}_v(U_j))$ is the mutual similarity between the two users in the two time intervals.

6.1.2 The HBM clustering algorithm

A clustering method, called hierarchical bisecting medoids (HBM) algorithm was developed to cluster users within time intervals based on the time-framed navigation similarity. One feature of this algorithm is that it avoids the common problem of requiring users to pre-specify on the number of clusters by using a hierarchical clustering technique. The algorithm combines features of the k-medoids and hierarchical clustering. Interested readers could refer to [9].

6.1.3 Mining association rules

The purpose of mining association rules is to find out which web pages are usually visited together in a session. Operated on the clusters of time-framed navigation sessions, the association rules discovered for each user session cluster will characterize the navigation patterns of specific user groups. As a result, these clustered association rules can serve as the knowledge models to predict the next navigation requests for future similar users.

6.2 Model application

6.2.1 User classification

The clustered association rules can serve as the knowledge base to give suggestion for future similar users. To achieve this purpose, a user classification method is needed to identify the cluster of navigation patterns to which the current user is most similar. Recall that each cluster of timed-framed sessions has a medoid, which is a frame of navigation sessions from some user. The medoid in some sense represents a typical user navigation pattern for users from that cluster. For a specific user, the session cluster to which the user is most similar can be selected by choosing the medoid to which the user's current behavior is most similar. The similarity computation is similar to the aforementioned similarity except that in this case we do not need to calculate the degree a medoid is similar to the user. That is, while the user's current behavior may be very similar to (or part of) that of the selected medoid, the medoid's behavior may have little similarity to the user's current behavior. As a result, we do not consider the mutual similarity between a medoid and the user, as its value may be very low due to the incompleteness of the user's current frame sessions.

A recommendation process is started right after a user has made his/her first request to a web site. After classifying user k as similar to a cluster, the association rules in the corresponding knowledge model of the cluster can be used to match the pages in the current session S_n of user k. Those rules matched with sufficient confidence (greater than a confidence threshold) will be fired, and the predicted items are added into the recommendation list in a sorted manner according to their decreasing confidence values. Furthermore, items that are suggested by more than one rules will be added to list only once with the highest confidence value.

However, it happens quite often that the current session of the user matches no association rules at all. So we need a recommendation mechanism that can provide reasonable suggestions when facing such a situation. Wang and Thao [9] proposed the following two mechanisms for this purpose.

6.2.2 The window-sliding method

This method uses a sliding window technique to control the number of session pages to be matched against the association rules. Let $S_n = [p_1, p_2, \ldots, p_k]$ be the user's current session. Initially, the window covers all pages in S_n, and hence all pages $(p_1, p_2, \ldots, p_k)$ in the current session are used to match against the association rules. If no matched association rules can be found, the window will slide one position to the right, leaving the pages $p_2, \ldots, p_k$ for rule matching. While the sliding actions will lose more and more information about the user's navigation behavior, it does preserve the most recent information as possible as it can. The sliding process will repeat until at least one rule is matched or the window coverage becomes empty. For the latter case, we say that the user cannot receive the recommendation service under his/her current navigation session.

6.2.3 The maximal-matching method

In contrast to using a sliding window method to preserve only the most recent session information for the matching work, the maximal-matching method preserves as much session information as possible for the matching work. This is achieved by finding all maximal subsets of the session pages that match successfully against the association rules. Given a set P of session pages, any subset M of P is called maximal if it matches at least one of the association rules, and no proper upper-set of M, which is also a subset of P, can find a matching rule. An efficient graph-based algorithm was implemented to find all the maximal-matching subsets of a page set given a set of association rules. To achieve this purpose, a lattice structure was used to store large itemsets discovered in the association-mining phase. Again, if no maximal-matching itemsets could be found, we say that the user cannot receive the recommendation service under his/her current navigation session.

6.3 Summary statements

Several factors have an impact on the performance of the recommendation method. They include: the time frame, the user classification method, the recommendation policies, the confidence threshold of recommendation, and the amount of training data. Historical navigation data was collected from three classes (classes A, B and C) of a virtual classroom course ('Expert System') for one semester. The experimental results show that the best average weighted precision rate is 0.6, average weighted recall rate is 0.7 and average service rate is 0.5, respectively. The results showed that the method is better in precision and recall rates than the conventional non-clustering one, and is comparable in the service coverage rate. The results also suggest that the recommendation method uses a shorter frame size such as a week for clustering user navigations and mining association rules, because a shorter frame size could track more flexibly the changes of users' traversal behavior. As to the recommendation policies, the results show that the maximal-matching policy is significantly better than the window-sliding one.

7 Concluding remarks

Data mining in e-learning is still in its nascent stage and needs much more research endeavor to make it of practical use for instructors and learners. There is more work worth further pursuit. For example, there are more effective mining tools for analyzing other kinds of learning activities, such as the behavior of 'thinking order' in a web-based online discussion context, and also those tools for helping teachers to explore the relationships between the various learning patterns and the learning outcomes [6, 29]. Teachers can analyze the stored student learning data to answer the questions such as 'what are the behavior characteristics of students tending to good learning outcomes?'

Besides, one limitation of current data association mining in e-learning is the inherent problem caused by the low supports of web page navigations, making it harder to build appropriate knowledge models. This is often solved by choosing

appropriately low support values used to mine the association rules, as is adopted in this research. Other association mining techniques [32] could be applied to avoid the low-support problem. Another area of work is to probe the effect of the knowledge model built by combining framed session clustering with mining sequential patterns.

A recent development has been in the use of learning objects which are cohesive pieces of learning material. With the emerging techniques of e-learning standards, instructors are able to disseminate their contents in the form of organized learning objects, instead of hyperlinked contents. Leaning objects represent learning experiences from instructors in a more compact representation form than the hyperlinked ones. Applying data mining on the usage analysis of learning objects is also an important research direction.

Finally, while data mining technology enables the provision of 'limited' personalized e-learning services, educational domain knowledge may be the key to make it really practical. For example, domain constraints might exist that could be used to confine the search space of data mining. Furthermore, content usage mining based on learning outcome will be able to give positive learning suggestions to learners. Pedagogical domain knowledge (such as educational thresholds, constraints, taxonomies and knowledge [33]), when integrated with the data mining technique, might be able to generate more flexible, efficient, contextualized and adapted learning environments. We need a mechanism of modeling and integrating the educational domain knowledge and data mining techniques to realize the dream of effective and efficient personalized e-learning.

References

[1] Arter, J.A., *Portfolios for Assessment and Instruction*, ERIC Clearinghouse on Counselling and Student Services: Greensboro, NC, 1995.

[2] Zaiane, O.R., Xin, M. & Han, J., Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. *Proc. of Advances in Digital Libraries Conf. (ADL'98)*, Santa Barbara, CA, pp. 19–29, 1998.

[3] Tang, C., Lau, R.W.H., Li, Q., Yin, H., Li, T. & Kilis, D., Personalized courseware construction based on web data mining. *Proc. of WISE Conf.*, pp. 204–211, 2000.

[4] Abramowicz, W., Kaczmarek, T. & Kowalkiewicz, M., Supporting topic map creation using data mining techniques. *Australian Journal of Information Systems*, **10**, pp. 63–78, 2004.

[5] Zaiane, O.R., Web usage mining for a better web-based learning environment. *Proc. of Conf. on Advanced Technology for Education*, pp. 60–64, 2001.

[6] Zaiane, O.R. & Luo, J., Towards evaluating learners' behaviour in a web-based distance learning environment. *Proc. of IEEE Int. Conf. on Advanced Learning Technologies (ICALT01)*, pp. 357–360, 2001.

[7] Mor, E. & Minguillon, J., E-learning personalization based on itineraries and long-term navigational behaviour. *Proc. WWW2004*, pp. 264–265, 2004.
[8] Wang, F.H. & Thao, S.M., A study on personalized web browsing recommendation based on data mining and collaborative filtering technology. *Proc. of National Computer Symp.*, Taiwan, pp. 18–25, 2003.
[9] Wang, F.-H. & Shao, H.-M., Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert Systems with Applications*, **27(3)**, pp. 365–377, 2004.
[10] Zaiane, O.R., Building a recommender agent for e-learning systems. *Proc. of the Int. Conf. on Computers for Education*, pp. 1203–1212, 2002.
[11] Barker, F.B., *Computer Managed Instruction: Theory and Practice*, Educational Technology Publications: Englewood Cliffs, NJ, pp. 4–10, 1979.
[12] Cooley, R., Mobasher, B. & Srivastava, J., Web mining: information and pattern discovery on the World Wide Web. *Proc. of IEEE Int. Conf. Tools with AI*, pp. 558–567, 1997.
[13] Chen, M.S., Han, J. & Yu, P.S., Data mining: an overview from a database perspective. *IEEE Trans. Knowledge and Data Engineering*, **8(6)**, pp. 866–883, 1996.
[14] Agrawal, R., Imielinski, T. & Swami, A., Mining association rules between sets of items in large databases. *Proc. of ACM SIGMOD*, pp. 207–216, 1993.
[15] Agrawal, R., & Srikant, R., Fast algorithm for mining association rules. *Proc. of The VLDB Conf.*, pp. 487–499, 1994.
[16] Han, J. & Kamber, M., Cluster analysis (Chapter 8). *Data Mining, Concepts and Techniques.* Morgan Kaufmann: San Francisco, CA: 2001.
[17] Gery, M. & Haddad, H., Evaluation of web usage mining approaches for user's next request prediction. *Proc. of the Fifth ACM int. workshop on Web Information and Data Management*, pp. 74–81, 2003.
[18] Eirinaki, M. & Vazirgiannis, M., Web mining for web personalization. *ACM Transactions on Internet Technology*, **3(1)**, pp. 1–27, 2003.
[19] Fu, X., Budzik, J. & Hammond, K.J., Mining navigation history for recommendation. *Proc. of the Fifth Int. Conf. on Intelligent User Interfaces*, pp. 106–112, 2000.
[20] Mobasher, B., Cooley, R. & Srivastava, J., Automatic personalization based on web usage mining. *Communications of the ACM*, **43(8)**, pp. 142–151, 2000.
[21] Mulvenna, M.D., Anand, S.S. & Buchner, A.G., Personalization on the net using Web mining. *Communications of the ACM*, **43(8)**, pp. 123–125, 2000.
[22] Lee, C.H., Kim, Y.H. & Rhee, P.K., Web personalization expert with combining collaborative filtering and association rule mining technique. *Expert Systems with Applications*, **21**, pp. 131–137, 2001.
[23] Riecken, D., Personalized views of personalization. *Communications of the ACM*, **43(8)**, pp. 27–28, 2000.
[24] Nichols, D.M., Implicit rating and filtering. *Proc. of the Fifth Workshop on Filtering and Collaborative Filtering*, pp. 31–36, 1997.

[25] Kashihara, A., Suzuki, R., Hasegawa, S. & Toyoda, J., A learner-centered navigation path planning in web-based learning. *Proc. of ICCE/ICCAI*, pp. 1385–1392, 2000.

[26] Kuo, C.-C., A data mining approach to auto-extraction of browsing structures of web materials. MD Thesis, Graduate School of Information Management, Ming Chuan University, Taiwan, 2001.

[27] Wang, F.-H., On analysis and modelling of student browsing behaviour in web-based asynchronous learning environments. *Lecture Notes in Computer Science*, **2436**, pp. 69–80, 2002.

[28] Breese, J.S., Heckerman, D. & Kadie, C., Empirical analysis of predictive algorithms for collaborative filtering. *Proc. of the Fourteenth Conf. on Uncertainty in Artificial Intelligence*, pp. 43–52, 1998.

[29] Liu, C.-C., Chen, G.-D., Ou, K.-L., Lee, C.-H. & Lu, C.-F., An instrument for on-line learning performance analysis by using decision tree technology on Web-based Portfolios. *Proc. of Int. Conf. on Computers in Education*, Japan, 1999.

[30] Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W.Y. & Wenyin, L., User intention modelling in web applications using data mining. *World Wide Web: Internet and Web Information Systems*, **5**, pp. 181–191, 2002.

[31] Lu, J., A personalized e-learning material recommender system. *Proc. of the 2nd Int. Conf. on Information Technology for Application (ICITA 2004)*, pp. 374–379, 2004.

[32] Wang, K., He, Y. & Han, J., Pushing support constraints into association rules mining. *IEEE Transactions on Knowledge and Data Engineering*, **15(3)**, pp. 642–658, 2003.

[33] Buchner, A.G. & Patterson, D., Personalized e-learning opportunities, call for a pedagogical knowledge model, *DEXA Workshops*, pp. 410–414, 2004.