Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Mining the Students' Learning Interest in Browsing Web-Streaming Lectures

Long Wang, Christoph Meinel
Hasso-Plattner-Institut, Unversity Potsdam
14482 Potsdam, Germany
long.wang@hpi.uni-potsdam.de, christoph.meinel@hpi.uni-potsdam.de

***Abstract*— Web-Streaming lectures overcome the space and time barriers between learning and teaching, but bring higher requirements on the learning feedback of students when they browse lectures. In this paper, we discover the students learning interest from their usage data in web-based learning environment by using multi data mining methods. The learning interests are expressed in six questions, which were asked by the teachers. We use simple statistics, associate rules mining, multi linear regression and similarity comparing to answer different questions. The usage data of online learners are heterogeneous, including HTTP server logs and REAL Helix Universal logs, and these heterogeneous usage data are transformed into students browsing profiles. We implement our work on our web-based learning environment: tele-TASK. The mined results help teachers to know their students clearly and adjust their teaching schedules efficiently.**

I. Introduction

One primary task of e-Learning systems is to supply e-Lectures for different online learners. The embedding of streaming video segments in distance learning modules has received great interest [11, 10, 9]. The basic feature of a web-based e-learning environment is to supply a number of different lectures for learners in semesters, and the fresh lectures are presented at the prominent position on the web, while the lectures for the past semesters can be accessed in archive pages. An ideal web-based e-learning environment is that: it displays not only the multimedia lectures, but also exercises, and even serves the final exams. But up to our acknowledge, till now there is no such perfect e-learning environment, and the web-based distance education systems, which serve mainly on supplying web-streaming lectures, are the currently most popular, reliable and efficient solutions. [9] stated that the web-streaming distance education offered today falls mainly into the one-way video and audio profile. The class video (along with the instructor audio) is typically recorded in a classroom studio-often filled with on-campus students- and posted on the class web site a few hours after the recording. The distance learners can then view the class video by streaming it from the class web site and interact with the instructor asynchronously, e.g. via e-mail or web-based discussion boards.

[10] studied the impact of the number of windows in web-streaming distance education video: one window video showing either the instructor or presentation slides/instructor writing pad; two-window distance education video, where one window displays the talking head of the instructor and the second window displays the presentationslides/writing pad; three-window distance education interface, where a live chat window is added.

A. Motivation

Such web-based e-learning systems facilitate teachers and learners greatly, but they can also lead to frustration for both of them, because the visual and aural cues (eye contact, body language, facial expression and voice tone) of online learners are missing compared to the education in face-to-face classroom. [19] reported a teachers' questionnaire to identify the needs of teachers to know their students and to make distance learning a less detached experience. They showed that the current e-learning environments have to be improved to satisfy teachers needs of tracking students in distance learning contexts.

[5] presented a tool aiming to track and analyze individual learner behavior during his interaction with e-learning environment. They suggested the further investigation on finding interesting patterns and navigation paths over set of single routers. [8] developed an education data mining tool which listens to the children when they read sentences and helps them learning how to read, but this application is not suitable for free and open web-based e-learning environment in high education, where the relationship between tutors and students is very loose and unstable.

[6] proved that different browsing strategies are used in different types of hypertext interfaces. Therefore it is necessary to enquire whether the type of hypertext architecture employed has any effect on the browsing strategies of individuals with different cognitive styles. In e-learning environments, different medium requires different way of evaluating student participation to ensure if the necessary knowledge or skills have been grasped during their learning.

To know the learning interest in web-based learning environments, which primarily deliver multimedia lectures, teachers can send a questionnaire directly to the learners, but they need the advanced mining tools to quantify the learning interest which include at least the followings questions:

1) *Are the online lectures welcomed by students?*
2) *Is there any difference between viewing the live broadcasting lectures and browsing lectures after they are recorded and edited?*

3) *Is there any preference on the different lectures in a course and preference on different pieces of one lecture?*
4) *Did the students view other lectures when they accessed one lecture?*
5) *Is there any relations between the exercise marks and the usage on lectures?*
6) *For the same named courses supplied for different years, is there any changes on the students' interest?*
7) *How often does student browse online lectures when they do their exercise?*
8) *How different between individual learning interest?*

The questions related with the learning interest could be raised more than those listed above. The educational mining tools including our methods can not solve all the above questions. Even in face-to-face classroom where direct questioning and answering are used, knowing students correctly is always the topic in pedagogy.

In this paper, we focus on answering the first six questions by different mining methods: general statistics, associate rules, multi linear regression and similarity comparing. The learning interest is mined from student learning profiles, which are transformed from heterogenous usage data. Our work is implemented based on our own web-based e-learning environment: tele-TASK, which is shown in the following subsection.

B. A Web-Based e-Learning Environment: tele-TASK

Tele-TASK (**T**eaching **A**nywhere **S**olution **K**its) [11] supplies a portable and powerful solution for distance education. From 2001 till 09.2006, tele-TASK has recorded over 500 different lectures and altogether more than 800 hours length recordings, and it has as well served in symposiums, conferences and other political events. All these lectures, multimedia recordings and other related materials are presented on web site: www.tele-task.de, which serves as the web-based distance learning platform. In this paper, we refer tele-TASK as its web site system, while not the lecture recording system. Students and interesting surfers can freely follow the live web-streaming lectures during semesters or the ongoing conferences.

All the web-streaming lectures and recordings are encoded in Real streaming format, and every lecture is embedded in a web page view for online learner to browse. The Figure 1 shows the snapshot of one web page view for a lecture.

The layout of one lecture page is divided into two parts: the left part and the right part. The left part is the outline of the whole course, which includes all the relevant lectures, and the text of each lecture name is linked to its streaming files. In most cases, one course includes several units (or chapters), in which there are several different lectures. The right part embeds the frame of the Real formatted streaming lecture. The frame of one streaming lecture is characterized by three fields: the top left field displays the "talking head" of the teacher synchronized with audio signal, the bottom left field writes the table of content (TOC) of the lecture, and each text line in TOC links directly to the right position in video that discusses the related knowledge. This field helps students to find their interesting knowledge easily and directly. The big right field shows the presentation slides/desktop or writing pad of the teacher.

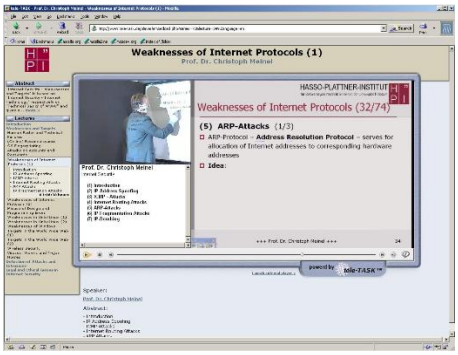

Fig. 1. One Lecture View on tele-TASK

Though the increasing access number on our web site and lots of recording requests convince us that tele-TASK helps to partly satisfy the great requirements in distance education, we do not know how are the web lectures used by students, and if there are some preference on different lectures. This is the motivation for this paper: we are trying to mine students' learning interest from their browsing behaviors, which could help us to know the learners and their learning interest, and to optimize the organization of the web site and the lectures.

Each multimedia lecture has definitive attributes after recording: **Name**, **Live Stream URL**, **Lecture Stream URL**, **Duration**, **Recorded Time**, **Table of Content**, **Course Name** and other attributes such as **lecturer** and **logo**.

II. Data Preparation for Mining Learning Interest on tele-TASK

A. Unifying Heterogeneous Learning Usage Data

This section explains how to filter and rebuild the right browsing data of online students. [15] stated that according to *www.kdnuggets.com/polls/2003/data_preparation.html*, two-thirds of data mining analysts consider that data cleaning and preparation consume more than 60 percent of total analysis time. We concentrate on this problem due to the extremely complexity and diversity of usage data in distance learning environment, which increase the difficulties to clean usage data.

The data on the web server side are the most common source of data for web mining, as they are easy to collect. Unfortunately, raw sever side data contain much noise and are usually incomplete. However, they contain useful data from which a well-designed data mining system can discover beneficial information. In most cases, a log entry is automatically added each time a request for a resource reaches the web server. Though this may reflect the actual use of the resource on a site, it does not record real behaviors like frequent backtracking or frequent reloading of the same resource when the resource is cached by the browser or a proxy.

Web-based e-learning environments usually supply heterogeneous learning materials including text, audio and video,

and store usage data in different format. In our case, tele-TASK web site records the normal surfing data on HTTP server in combined log format, and at the same time stores the usage data on streaming lectures which run on real helix server. Every web-streaming lecture is embedded in one page view, which means that when a student clicks the lecture link, two totally different usage log entries will be written on two different servers in different log format. It is planned we will develop our own logging mechanisms to record the required usage information in one unified format.

The free accessing on tele-TASK site brings the complexity and difficulty to recognize students' learning interest. We separate all the users into three kinds: the students, the instructors and others such as crawlers, robots and irrelevant visitors. In order to mine the students learning interest, it is necessary to filter out the last two kinds of usage data:

- *Instructors and administrators have the constant IP addresses, so we can easily remove the requests sent from these IP addresses.*
- *Recognizing web robot sessions was discussed in [14]. In web-based educational environments, some web robots are identified by IP addresses and user agents of web clients. Others can be recognized by their crawling patterns: firstly, one web robot spends little time on fetching or doing actions on streaming lectures; secondly, there is no clear preference on different lecture pages, which can be censored by the time interval between every two successively fetched pages and by the ratio of the number of fetched pages compared with the total number of pages in the learning environment.*

HTTP server logs related with web usage mining were widely discussed in [3, 1, 16], the useful usage information for each log entry includes: ***IP address***, ***request time***, ***request file***, ***user agent*** and ***referee link***. To understand the web usage patterns, it is required to know their accessing objects. For mining students' learning patterns, the accessing objects such as lecture descriptions and content schedule have to be retrieved from HTTP server logs. A page view shown in web browser usually generates several log entries including different request files in server logs. HTTP server logs record only the name or the path for request files, without the content and semantic information for them. Different request lines may link to the same content, which is one of the most popular problems in WWW. Dynamical web sites, which are characterized by generating the content of page views based on the input or the client configurations of the visitors, bring much more difficulties to retrieve the accessing objects from request lines in server logs. This task would be impossible if there is no extra supports, such as the site map, functionalities between site structure and physical file systems or logging mechanism that records the detailed interactions between server and visitors.

Tele-TASK web site runs on PHP server, and generates page views by fetching series, course and lecture descriptions from MySQL database parameterized by client session ID.

```
IP_address - - [timestamp] "GET filename protocol/version" HTTP_status_code
bytes_sent [client_info] [client_ID] [client_stats_results] file_size file_time
sent_time resends failed_resends [stream_components] [start_time] server_address
```

Fig. 2. Helix Universal Access Log Format in Logging Style 3

For the lecture page view, besides the text information of the lecture from MySQL, it invokes fetching lecture video from Real Helix server. In cases, one single streaming lecture is involved in different courses, and this happens when some courses named the same title for different semesters have some chapters sharing the same content organizations. This is also one of the conveniences that web-based teaching brings. Such flexibilities cause the variant URLs for the single same accessing object in database. To retrieve the unique accessing objects from the request lines in tele-TASK HTTP server logs, we generalize the URL generating rules written in an XML file.

Real Helix server supplies 6 logging styles, which collect different information of access requests. For example, the format for logging style 3 is shown in Figure 2. We have showed that one lecture view is divided into three fields: "talk header" video field, desktop field and table of content field. The former two fields play two *.rm files, while the content field displays one *.rt file. Different pieces of RM files can play in a time sequence and comprise one "talk header" video or one desktop video.

B. Model Student Browsing Profile

The process of browsing multimedia lectures and other relative information can be seen as the online learning session. This learning session can be depicted as: one student views one learning object, if he finds that the knowledge from the content abstract is very familiar to him, he stops viewing this object and goes on finding other learning objects or just leaves web site; if he finds the learning object is his interesting target, he goes deeply viewing this object; if he finds he can not understand some pieces of his learning object in one view, he repeats viewing these pieces.

The learning objects presented in page views are classified in two kinds: with and without embedded multimedia lectures. The later are pages dedicated on the outline descriptions for courses, colloquium or other topic units, and they often link to pages with embedded lectures.

The above preprocessing helps to filter and translate the heterogeneous raw usage data into a set of browsing events. Each browsing event is represented as: ***session***, ***student***, ***learning object***, ***type***, ***start time***, ***duration*** and ***operation***. The first five parameters can be computed by the method discussed in the former section.

Now we explain how to compute the ***duration*** of one student spent on one learning object during one learning session:

- *if the learning object is multimedia lecture, the* ***duration*** *is computed by* ***timestamp*** *subtracted by* ***start_time****, both are recorded in real helix server log styled 3;*
- *if the learning object is normal page view like course or series description, the* ***duration*** *is defined as the gap*

between the time stamp of this learning object and that of the just next object recorded in the same learning session, and it is set to 0 if this learning object is the last one.

We assume during this learning session every student kept on sitting in front of his computer and concentrated on learning, though online learners require much more maturity, more self-motivation and self-discipline than those in traditional classrooms [17].

Individual learning ***operation*** can not be directly measured from usage logs. As shown in former section, the table of content field in online lecture view composes several sub headlines linked to the right positions within multimedia lectures, which facilitates online learners to reach directly to the right interesting piece, and the slide bar at the bottom of media lecture helps learners select or repeat some piece of lectures as well. When one online student clicks the hyper links in content table, jumps over some piece or repeats some piece, stops and resumes the lecture, the helix server will stop the current ongoing lecture and reload the right piece in **RM** files, and all these transfers dedicated to the same learner and the same lecture are recorded in helix server logs. The time-stamp gap between every two transferring depends on the actions that the learner made:

- *if the actions are clicking hyper links, jumping over and repeating, the gap will last several seconds depending on the network overloading;*
- *if the actions are stopping and resuming, the gap will be decided by the two clicks from the client learner plus the transferring delay.*

Such usage data help us to assess the operations of one online learner on multimedia lectures. Assuming that the right usage data have been separated for one learner from helix server logs and cut for different learning sessions from the usage data of this learner (the detailed techniques were widely discussed in [3, 1]), the operations of this learner can be estimated from the number of records on the same lecture within one learning session. Moreover, the learning duration on the one lecture within a learning session is the sum of all the durations extracted from the records on the same lecture.

However, there is an exception in computing learning operations on lectures: if there exist several records on the same live broadcasting lecture within a learning session, which means that the learner could not jump over or repeat some piece and the reason of multi recordings was mainly the network overload or just client's clicking stop button, we induce such records within one learning session to one record and the learning operation is concluded as 1. Table 1 shows one piece of student browsing profiles on tele-TASK:

TABLE I
EXAMPLE OF STUDENT BROWSING PROFILE

Session	Student	Learning Object	Type	Start Time	Duration	Operation
...	...	...	...	...	...	...
547	736	www12	0	31/May/2006:08:48:53	00:00:20	1
548	737	www12	1	19/Jun/2006:19:27:55	00:56:33	2
548	737	TI_08	1	19/Jun/2006:20:31:55	00:21:42	6
...	...	...	...	...	...	...

Within the same learning duration, one learner with more operations displays more interest than that with few operations: the former finds clear and concrete learning object in the lecture, while the later is possibly a fresh learner on this lecture. But we can not guarantee that some operations within few learning sessions were due to the reloading of network or real helix server.

III. METHODS ON MINING LEARNING INTEREST ON WEB-STREAMING LECTURES

The current distance education environments work mainly as the secondary supplement for the conventional education. In conventional education, the teachers and education supervisors can ask the students face-to-face or use anonymous questionnaires to judge if their lectures taught in classroom are welcomed or not. In other web services, such as online shopping or e-communities, it is relatively easy to evaluate the success of their services by the changing number of online bills or the number of members.

But in distance education environments, especially in free and open environments only supplying web-based streaming lectures, it becomes much more difficult to evaluate the success of online lectures. Surely, the changing number of accessing is an important indicator, but it is not enough, just like [12] said "using hits and page views to judge site success is like evaluating a musical performance by its volume".

Before answering the six questions showing the multi facets of online learning interests, it is necessary to define some parameters on learning interest, which are shown in Table 2.

TABLE II
DIFFERENT PARAMETERS FOR LEARNING INTEREST

Parameter	Meaning
NS_C	Number of students that register to choose the course C
$NC_{S,l}$	Number of students that attend l in classroom
$NA_{S,l}$	Number of accessing live streaming version of l
$NP_{S,l}$	Number of accessing post edited version of l
$ND_{S,l}$	Average time duration of viewing l
$NO_{S,l}$	Average Number of Operations of viewing l

A. Question 1: Are the online lectures welcomed by students?

The ever increasing or decreasing of $NA_{S,l}$ and $NP_{S,l}$ compared to that on $NC_{S,l}$ shows if the lectures are welcomed by online learners or not. $NA_{S,l}$ and $NP_{S,l}$ can be directly computed from usage logs, and $NC_{S,l}$ can be gotten by asking the teachers or their teaching assistants.

There exist the noise in computing $NA_{S,l}$ and $NP_{S,l}$: a part of online students, who viewed the lectures, are not included in NS_C, and their usage records are mixed in those of the students who enrolled this course and it is even impossible to assess the percentage of such noise. This is the drawback that the total open and unauthorized web based teaching systems bring.

B. Question 2: Is there any difference between viewing the live broadcasting lectures and browsing lectures after they are recorded and edited?

The comparison between $NA_{S,l}$ and $NP_{S,l}$ tells the the preference between viewing the live broadcasting and the edited lectures. The time duration of the live broadcasting of one lecture is decided by the length of lecture's recording, and it is usually between 60 minutes and 90 minutes. It can be predicted that $NA_{S,l}$ is always less than $NP_{S,l}$, but we can use the changes of $NP_{S,l}$ based on the day, week or month to find the detailed difference between $NA_{S,l}$ and $NP_{S,l}$.

C. Is there any preference on the different lectures in a course and preference on different pieces of one lecture?

The preference on different lectures can be computed by comparing their $NA_{S,l}$, $NP_{S,l}$, $ND_{S,l}$ and $NO_{S,l}$. One lecture with bigger $NA_{S,l}$ and $NP_{S,l}$ shows much more acceptance than that with smaller $NA_{S,l}$ and $NP_{S,l}$. Further, one with bigger $ND_{S,l}$ and $NO_{S,l}$ tells that students would like to spend more efforts on it than that with smaller $ND_{S,l}$ and $NO_{S,l}$ if there is no big difference between two lectures on $NA_{S,l}$ and $NP_{S,l}$.

$ND_{S,l}$ can be computed as follows: $ND_{S,l} = \frac{\sum duration}{NA_{S,l}+NP_{S,l}}$; where ***duration*** is the time that one online learner spent on this lecture during one learning session and can be directly fetched from student browsing profiles. $NO_{S,l}$ can be computed as: $NO_{S,l} = \frac{\sum Operatioin}{NA_{S,l}+NP_{S,l}}$.

D. Question 4: Did the students view other lectures when they accessed one lecture?

Answering this question can be formulated as mining the frequent lecture sub sets of the lecture set L over the set of student browsing profiles F. Mining such relations is a typical example of mining association rules or frequent item sets [3, 4, 15]. The implicit relations among different online lectures could help teachers to know if they need to combine some lectures or add some contents from other courses.

We simplify a learning session p of a student s on some lectures as: $p_s = \{l_1...l_k\}$, where $l_i \in L$ and $s \in F$. Transformed from the set of student browsing profiles F, the set including all the learning sessions are named as P. From P, we try to mine the relations each of which is formed as $r = \{l'_1...l'_t\} : Supp_r$, where $l'_i \in L$ and $Supp_r$ is the number of sessions that viewed all the lectures in r. The methods to mine association rules or frequent item sets have been widely discussed. We used the mining method referred in [4], which integrates all the leaning sessions into a highly compressed extended prefix-tree structure called frequent pattern tree stored in memory, and the complete frequent item sets can be mined from this tree structure without candidate generation.

E. Question 5: Is there any relation between the exercise marks and the usage on lectures?

Such kind of relations help the teacher to know if the mastery levels of students are decided by their viewing lectures, and if the problem is due to his lectures when there were lower marks with higher accessing numbers. The exercises are always delivered to the students after recording the necessary lectures, and students are asked to find the answers with the help of lectures within few days.

The exercise marks are always given in a number format or the rankings. The average mark of one exercise sheet from all the students is the general indicator for the mastery level of the students. On the other hand, the usage on every online lecture has multi attributes. As explained in former sections, $NA_{S,l}$, $NP_{S,l}$, $ND_{S,l}$ and $NO_{S,l}$ are the necessary parameters to show the online usage on lectures. One indicator depicting the general usage on one lecture has to be computed from these four different facets of usage. Such computation is generalized as the problem of multi input single output problem solved by multi linear regression. We use the ***usage score*** of a set of students S on one online lecture l to name this indicator, and it is computed as:

$$US_{S,l} = \alpha \times NA'_{S,l} + \beta \times NP'_{S,l} + \gamma \times ND'_{S,l} + \delta \times NO'_{S,l} + \theta. \quad (1)$$

where $\alpha + \beta + \gamma + \delta + \theta = 1$, and the values of these five coefficients are assigned based on the statistical observations or the expert experiences. Before computing the usage score of S on L, four variables have to be normalized.

The relation between the exercise marks and the usage on lectures is interpreted by comparing the higher or lower of average exercise mark with that of usage score on the lectures.

F. Question 6: For the same named courses supplied for different years, is there any changes on the students' interest?

For two courses served in different years with the same name, the changes of the students' interest in this year compared to the other year could tell the trends of learning interest and the teacher can compress or enlarge some contents based on this trends. Measuring the change of learning interest is the contrary to computing the similarity of learning interest on two courses: the similarity less than the defined threshold means that the learning interest is changed from the last year to this year, and the learning interest can be seen stable if the similarity is larger than the threshold. The threshold is defined by experienced teachers.

One course is usually made different from others by its relatively clear conceptual intention and extension, and usually characterized by a set of related knowledge elements. These knowledge elements will be delivered to the learners in a suitable sequence in several lectures. The lectures having closer relations always form the subdivisions of the course, and these subdivisions are called chapters (units). The relations among course, chapters, lectures and knowledge elements can be formulated in a reverse tree structure: the top (root) layer is the course and the bottom (leave) layer comprises all the knowledge elements, the interior two layers are chapters and lectures. The structure of one course is shown in Figure 3.

The learning interest on a course C over a set of students S is decided by the learning interests on all the lectures included in C:

$$Interst(S, C) = \{Interst(S, l_1), Interest(S, l_2)...Interst(S, l_n)\}, \quad (2)$$

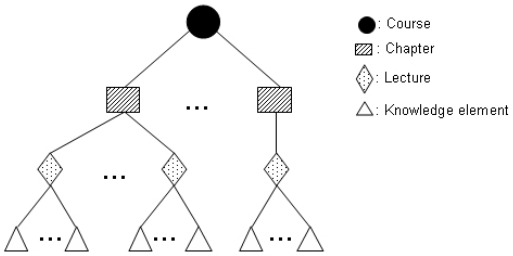

Fig. 3. Knowledge Structure of a Course

where $l_i \in C1 \leq$ and $i \leq n$.

Because the changes of learning interest on the whole course hide the changes on single lecture, and the changes on single lecture can not depict the changes on the whole course, we investigate the changes of learning interest from three levels: course level, chapter(unit) level and lecture level.

1) From Course Level: From the course level, the simple way to measure the changes of learning interest is to compare the two summations of the learning interest on each lecture included in two separate courses:

$$Change(I_{S_1,C_1}, I_{S_2,C_2}) = \frac{\sum Interest(S_2, l_{2i}) - \sum Interest(S_1, l_{1j})}{\sum Interest(S_1, l_{1j})}, \quad (3)$$

where $l_{1i} \in C_1$ and $l_{2j} \in C_2$. The learning interest has five parameters discussed before as: NA, NP, ND, NO and US.

2) From Lecture Level: The changes of learning interest on two lectures are not only decided by the usage shifting, but also affected by the difference of the lectures. To measure the changes of learning interest on lectures, it is necessary to integrate the changes of the lectures. Given one lecture l_1 in one course and its corresponding lecture l_2 in another course, the change of learning interest from l_1 to l_2 is computed as follows:

$$Change(I_{S_1,l_1}, I_{S_2,l_2}) = \frac{I_{S_2,l_2} - I_{S_1,l_1}}{I_{S_1,l_1}} \times Sim(l_1, l_2). \quad (4)$$

We judge the similarity between lectures l_1 and l_2 by comparing their knowledge elements, which can be extracted directly from the TOC field of the lecture view. Assuming that all the knowledge elements belonging to one single lecture play the same weight on characterizing this lecture, and K_1 and K_2 are the sets of the knowledge elements of l_1 and l_2 separately, the similarity between two lectures can is computed as:

$$Sim(l_1, l_2) = \frac{K_1 \cap K_2}{K_1 \cup K_2}. \quad (5)$$

One lecture usually includes 5∼15 knowledge elements, so the cost on computing the similarity between two lectures is very few.

3) From Chapter Level: Seen from Figure 3, one chapter from a course is a sub tree formed by lectures as interior nodes and knowledge elements as leaf nodes. The changes between two chapters can be measured by the edit distance [2] between two chapter. Similar to [18], we define basic tree edit operations on the chapter:

- $Insert(x, y)$: *insert a node x as a leaf node of node y.*
- $Delete(x, y)$: *delete a leaf node x from node y.*
- $Update(x, l)$: *update a node x in T with the new label l resulting that T is identical to T', which means that T' is identical to T except that the label of x is l.*

The edit operation $Update(x, l)$ applied on a leaf node can also be realized by the combination of two edit operations $Delete(x, y)$ and $Insert(l, y)$, where y is the father node of x and l.

Based on the edit operations, an *edit script* is defined as a sequence of basic edit operations that transform one tree to another [2]. The $cost$ of an edit script is defined to be the sum of the costs of its basic edit operations. The *edit distance* between trees T_1 and T_2 is to find a minimum *cost edit script* that transforms T_1 to T_2. Considering the intuitive and natural way to assign identical costs to $insertion$, $deletion$ and $update$ operations ($Cost_{Ins} = Cost_{Del} = Cost_{Upd} = 1$), so the *edit distance* between trees T_1 and T_2 is defined as the number of basic edit operations in the edit script, and we use $Dist(T_1, T_2)$ to name this edit distance. So the similarity between trees T_1 and T_2 is computed as:

$$Sim(T_1, T_2) = \frac{max(Dist(, T_1), Dist(, T_2)) - Dist(T_1, T_2)}{max(Dist(, T_1), Dist(, T_2))}. \quad (6)$$

where $Dist(, T_1)$ and $Dist(, T_2)$ are the edit distances to build T_1 and T_2 from empty separately. The algorithm for computing edit distance can be refereed in [2].

The learning interest of a set of student S on one chapter T is computed as the sum of learning interest of S on all the lectures in T: $I_{S,T} = \sum Interest(S, l_i)$, where $l_i \in T$. So the changes from I_{S_1,T_1} to I_{S_2,T_2} is computed as:

$$Change(I_{S_1,T_1}, I_{S_2,T_2}) = \frac{I_{S_2,T_2} - I_{S_1,T_1}}{I_{S_1,T_1}} \times Sim(T_1, T_2). \quad (7)$$

IV. Analysis of Mining Results on tele-TASK

We implement our mining methods in our web-based learning environment: tele-TASK, and the usage data include: access logs from HTTP server, access logs from helix universal server and the exercises marks of the learners. Both kind of access logs are taken from one teaching semester 01.04.2006∼31.07.2006 (LOG_I), and record the learning data on three courses: Technic Basis of WWW(WWW), Theory Information(TI) and Operation System Architecture(BSA). The course of WWW includes 26 lectures, TI comprises 15 lectures and BSA has 21 lectures. To compute the changes of learning interest on the course WWW, we take the access logs during 01.04.2005∼31.07.2005(LOG_{II}), from which the learning interest on WWW will be compared to that from LOG_I. However, the exercise marks were only available for the course WWW in the semester 01.04.2006∼31.07.2006.

A. General Information on Learning Feedback

From LOG_I, we filtered out 2991 learning session (LP_s), 2346 of which are the sessions of viewing the edited lectures (LP_e), while the others are those that only viewed the live broadcasting. From LP_e, we further found that 887 learning sessions are the learning on WWW, 902 on TI and 817 on

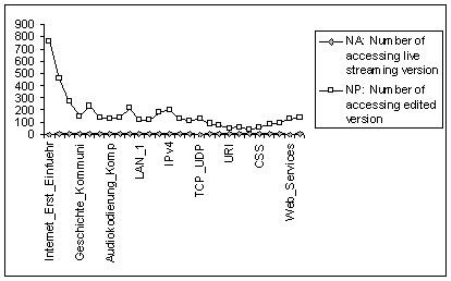

Fig. 4. NA and NP on the lectures in WWW

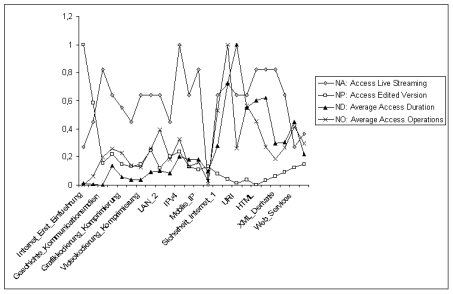

Fig. 5. Usage on lectures of WWW

BSA. From LOG_{II}, we extracted out 448 learning sessions on WWW.

TABLE III
MAXIMUM, MINIMUM AND AVERAGE USAGE ON WWW LECTURES

	NA	NP	ND	NO
Maximum	11 (IPv4)	767 (Erst Einfuehrung)	00:30:12 (Sicherheit Internet 2)	8.11 (Sicherheit Internet 2)
Minimum	0 (TCP UDP)	34 (HTML)	00:00:19 (Erst Einfuehrung)	2.68 (Erst Einfuehrung)
Average	7	164	00:08:37	4.2

B. Answering Different Questions

We take the example of course WWW in LOG_1 to answer the question 1, 2 and 3. Altogether 21 students have chosen this course. The number of students that attended the lecture is about 6 on average. Figure 4 records the NA and NP on the lectures of WWW. The numbers of accessing edited version and live streaming version compared to the number of student choosing this course show that the online lectures are welcomed by the students. NA is always close to NC, and on some lectures even larger. The explosive number of NP compared to NA tells that the students usually viewed the edited lectures after they viewed the live streaming or attended the lecture in classroom. We can also draw that students usually viewed the edited lectures if they missed the live streaming version, and this can be concluded from the lectures "TCP/UDP" on which NA is 0 while NP is large compared to other lectures.

Figure 5 shows the different facets of usage on lectures of WWW. From this figure, we find that the lecture of "Hypertext Markup Language" has the lowest access number, and

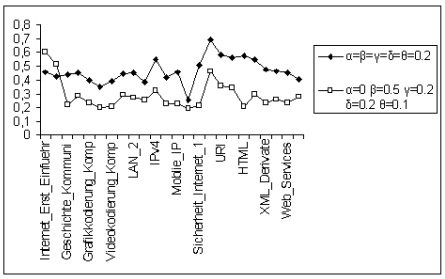

Fig. 6. Usage Score on WWW lectures

the lectures of "Uniform Resource Identifier" and "Hypertext Transfer Protocol" have also the lower access numbers. On the other hand, even the lectures of "Web Services" and "Semantic Web" were presented at the end of the semester, but they still attracted many visits. This shows that the students have already been familiar to the basic knowledge such as "HTML" and "URI" before they choose this course. We can also draw this conclusion from the lower learning interests on these three lectures.

Based on the formula (1), we compute the usage score on different lectures, and compare to the exercise marks. Figure 6 displays the results of two assignments on different coefficients: $\{\ \alpha = \beta = \gamma = \delta = \theta = 0.2\}$ and $\{\ \alpha$=0, β=0.5, γ=0.2, δ=0.2, θ=0.1$\}$. The second one thinks that the number of accessing edited lectures plays much important than that of accessing live streaming.

To find if there exist some relations between the lectures during learning sessions, we set the threshold of the support number 1% to mine the sub sets of frequent lectures. We find that the first two lectures of WWW were always viewed together, this happened as well in courses TI and BSA. In WWW, the lectures belonging to the same knowledge category were viewed together, such as "LAN(I)" and "LAN(II)", but this is not true to the lectures "Sichereit im Internet(I)" and "Sicherheit im Internet(II)". We find that there is no relation among lectures belonging to different courses, and the low threshold and few mined relations suggest us that most of the online learners have clear and singular learning objects during one learning session.

The changes of usage on WWW from 2005 to 2006 is shown in Figure 7. Due to the popular acceptance by students on e-learning and efficient arrangement of teachers on teleteaching materials, the lectures in 2006 attract much more learning interest than those in 2005, no matter from any aspects of usage. The accessing number on the edited lectures raised explosively, and web students spent much more time than before, and their interactivities with the lectures become more active as well.

The usage change in Figure 7 does not show the effect of the changes on the lectures. The changes of learning interest integrating both changes on usage and lectures are shown in Figure 8. Form Figure 8, we can find its difference compared to Figure 7. Though explosively increasing of learning interest

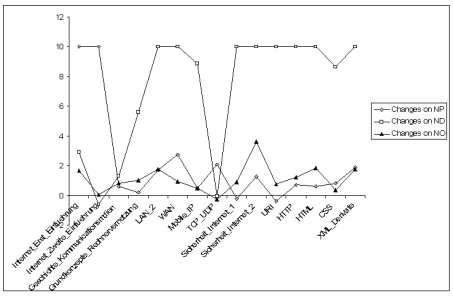

Fig. 7. Usage Changes on WWW from 2005 to 2006

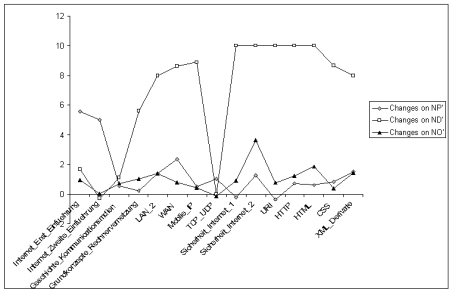

Fig. 8. Learning Interest Changes on WWW from 2005 to 2006

on most of lectures, the decreasing of that on "TCP/UDP", "URI" and "Sicherheit Internet 1" helps the teachers to think about if they know correctly the students' mastery levels.

Based on these facts, we can adjust the course WWW in the future: to delete or compress the lectures on "HTTP", "HTML" and "URI", while to enlarge the lectures about "Web Services", "Semantic Web" or other knowledge on the frontier of WWW.

V. Conclusions and Future Work

In this paper, we have investigated mining learning interest in browsing web-streaming lectures. We explain the learning interest by raising 6 questions from the teacher's view, and different questions can be formulated into different mining targets requiring different mining methods. The complexity of e-learning environment requires extra work on cleaning learning data and transforming cleaned learning data into learning profiles. The learning interest of a set of students on a lecture is multi linear regressed from four learning attributes of this lecture. We mined the changes of learning interest on the same titled courses served in different semester from three different levels. Some interesting information are mined from the learning on tele-TAK: we are confirmed that the online learners usually have clear learning objects, but we are surprised that some lectures were not welcomed by the learners, while some are greatly required by the learners.

The average time of a student spending on a lecture is about 10 minutes, while the normal length of a lecture is about 90 minutes. This gives us the hints to segment the whole lecture video into small pieces and to organize them with a searchable semantic network, which could help students finding suitable and related knowledge during learning. On the other hand, the relations between usage on lectures and the exercises marks are not so clear. The reason is that the organization of exercises and lectures can not be mirrored to each other and the usage is dedicated to the lectures, not to the knowledge within the lectures. This pushes us to develop our own learning logging mechanism recording the detailed interactivities between learning and lectures.

Acknowledgment

We thank Dirk Cordel to collect the students' marks of exercise, middle and final examinations.

References

[1] B. Berendt, M. Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, 9: 56-75, 2000.

[2] S. Chawathe. Comparing Hierarchical Data in Extended Memory. In Proceedings of the Int. Conf. VLDB, 1999.

[3] R. Cooley, B. Mobasher and J. Srivastava. Data preparation for mining world wide web browsing patterns. Journal of Knowledge and Information Systems, 1(1), 1999.

[4] J.W. Han, J. Pei and Y.W. Yin. Mining Frequent Patterns without Candidate Generation. In Proceedings of Int. Conf. ACM SIGMOD, 2000.

[5] J. Hardy, M. Antonioletti and S. Bates. e-Learer Tracking: Tools for Discovering Learner Behaviour. In Proceedings of Int. Conf. IASTED Web-Based Education, 2004.

[6] R. McAlesse. Navigation and browsing in hypertext. In Hypertext: theory into practice, 1999.

[7] A. Merceron and K. Yacef. TADA-Ed for Education Data Mining. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 7(1), 2005.

[8] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea and C. Heiner. An Educational Data Mining Tool to Browse Tutor-Student Interactions: Time Will Tell!. In Proceedings of Int. Conf. AAAI, 2005.

[9] J. Reisslein, P. Seeling and M. Reisslein. Video in distance education: ITFS vs. web-streaming: Evaluation of student attidues. Internet and Higher Education, 8 (25-44) 2005.

[10] P. A. Reynolds, R. Mason. On-line video media for continuing professional development in dentistry. Computer and Education, 39(1), 65-98, 2002

[11] V. Schillings and C. Meinel. tele-TASK - Teleteaching Anywhere Solution Kit. In Proceedings of Int. Conf. ACM SIGUCCS, 2002.

[12] E. Schmitt, H. Manning, Y. Paul and J. Tong. Measuring web success. Forrester Report, November 1999.

[13] M. Spiliopoulou, B. Mobasher, B. Berendt and M. Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis. INFORMS Journal on Computing, 171-190, 2003.

[14] P. N. Tan and V. Kumar. Discovery of Web Robot Sessions Based on their Navigational Patterns. Data Mining and Knowledge Discovery, 6, 9-35, 2002.

[15] D. Tanasa and B. Trousse. Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Transactions on Intelligent System, 59-65, March/April 2004.

[16] L. Wang C. Meinel. Discovering Characteristic Individual Accessing Behaviors in Web Environment. In Proceedings of Int. Conf. RSFDGrC, 2005.

[17] D. Zhang, J. L. Zhao, L. Zhou and J. F. Nunamaker. Can e-Learning Replace Classroom Learning? Communications of The ACM, Vol. 47, No. 5, 2004.

[18] Q. Zhao, S. Bhowmick, M. Mohania and Y. Kambayashi. Discovering Frequently Changing Structures from Historical Structural Deltas of Unordered XML. In Proceedings of the ACM Int. Conf. CIKM'04.

[19] C. Zinn and O. Scheuer. Getting to Know your Student in Distance Learning Contexts. In Proceedings of European Conf. TEL, 2006.