

Detecting the Changes of Web Students' Learning Interest

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

Abstract

In this paper, we discover the changes of students' learning interest from their usage data in web-based learning environment. Due to the effects on each other of the changes in web students and web lectures, we seek a method that integrates the changes in both sides to measure the changes of learning interest. 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.

1 Introduction

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. A teachers' questionnaire reported in [10] showed that the current e-learning environments have to be improved to satisfy teachers' needs of tracking students in the contexts of distance learning. Among these needs, one prominent is to know the changes of learning interest on the same titled courses serving in different years. Though some clever web mining tools can mine the usage of online visitors on their web services [2, 4, 6, 7], changes of usage interest were few discussed in details [1,5,8].

Tele-TASK (Teaching Anywhere Solution Kits) [9] supplies a portable and powerful solution for distance education. From 2001 till 09.2006, tele-TASK has recorded over 500 different lectures and more than 800 hours length recordings, and it has as well served in symposiums, conferences and 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. 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 all the lectures used by students. This is one big motivation for

this paper.

We describe the learning interest of a set of web students on a learning object during one period of time in a gram comprised by *number of accessing live streaming version*, *number of accessing posted version*, *average usage time length* and *average usage operation*. Based on this trigram, we further compute the usage score which synchronizes the effects from different usage aspects. We use the *structural similarity* to measure the changes between two courses serving in different semesters, and propose an approach to compute structural similarity on different levels.

The rest of this paper is organized as follows. Section 2 gives the necessary preliminary statements and definitions and discusses the methods for measuring the changes of learning interest. Section 3 presents our experiment results on tele-TASK. Finally we give our conclusion in section 4.

2 Methodology for Measuring the Changes of Learning Interest

One web course has its clear conceptual intensions and extensions, and it is characterized by a set of knowledge elements. These knowledge elements will be delivered to the web students in the form of web-streaming lectures in a suitable sequence. Further, the lectures belonging to the same course are organized in several non-intersected sub sets: *units* or *chapters*. The knowledge elements k , lectures l , units u and course U are organized in a tree hierarchy structure, and we simplify this tree structure into four layers: *course layer*, *unit layer*, *lecture layer* and *knowledge element layer*.

Similar to [3], we define the following *basic edit operations* for computing the changes between two courses(units, lectures):

- *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 .

Based on the basic edit operations, a *structural edit script* is a sequence of basic edit operations that convert one structure to another. Here the *structure* has different levels: course level, chapter level and lecture level.

Definition 1 (Structural Distance) Let l_1 and l_2 be two lectures, structural distance SD_{l_1, l_2} is the number of basic edit operations in the structural edit script that can change l_1 to l_2 .

Based on the *structural distance* from l_1 to l_2 , we compute the *similarity measure* between l_1 and l_2 as:

$$SM(l_1, l_2) = \frac{\max(SD(\emptyset, l_1), SD(\emptyset, l_2)) - SD(l_1, l_2)}{\max(SD(\emptyset, l_1), SD(\emptyset, l_2))}. \quad (1)$$

where $SD(\emptyset, l_i)$ ($i \in \{1, 2\}$) is the *structure distance* of building the entire l_i from empty based on the basic edit operations.

Given a web-streaming lecture l , we use $UG_{S, l} = (NA_{S, l}, NP_{S, l}, ND_{S, l}, NO_{S, l})$ to name the *usage* of a set of web students S on l during one time period T , which are explained in table 1.

Table 1. Explanations of $NA_{S, l}$, $NP_{S, l}$, $ND_{S, l}$ and $NO_{S, l}$

Variables	Meaning
$NA_{S, l}$	Number of accessing live broadcasting version of l
$NP_{S, l}$	Number of accessing posted version of l
$ND_{S, l}$	Average usage time length on l
$NO_{S, l}$	Average usage operation on l

Definition 2 (Usage Score) Let $UG_{S, l} = (Na_{S, l}, Np_{S, l}, Nd_{S, l}, No_{S, l})$ be the usage of one web student set S on lecture l during one period time, the usage score of $UG_{S, l}$ is defined as:

$$US_{S, l} = \alpha \times NA'_{S, l} + \beta \times NP'_{S, l} + \gamma \times ND'_{S, l} + \delta \times NO'_{S, l} + \theta. \quad (2)$$

where $\alpha + \beta + \gamma + \delta + \theta = 1$.

The *usagescore* is an aggregation of various usage aspects in the online courses, and shows the general learning preference from a set of online students. Each aspect is assigned a percentage weight. This assigned weight is based on the subjective importance determined by the teacher.

Definition 3 (Changes of Usage) Let UG_{S_1, l_1} and UG_{S_2, l_2} be two usages, changes of usage $CH(UG_{S_1, l_1}, UG_{S_2, l_2})$ are the increasing or decreasing of usage parameters and usage score in one time period compared to another time period.

To take into account content changes in detecting changes of learning interest, we propose to vary changes of usage based on the following equation:

$$CHI(S_1, l_1, S_2, l_2) = SM(l_1, l_2) \times CH(UG_{S_1, l_1}, UG_{S_2, l_2}). \quad (3)$$

Given two sets of usages on the same titled courses serving in different semesters, the problem of detecting the changes of learning interest is to find the learning objects from different levels (course, chapter, lecture) on which the learning interests are changed beyond the predefined thresholds integrated with the change of learning objects. For example, we set $\tau_u = 0.05$ for US and $\tau_n = 0.08$ for NA , NP , ND and NO to mine the lectures on which the change on US is out of $[-5\%, 5\%]$ and change on **any** of NA , NP , ND and NO is out of $[-8\%, 8\%]$.

The algorithm to compute *structural distance* between two trees can be referenced in [3]. But our aim is to compute the similarity on different levels for every two course structures, and we concentrate on the *structural difference*. Though it is easily proved from the definitions of structural similarity that computing similarity between two trees in one level is symmetric, the similarity comparisons on all the levels are asymmetric. Computing the changes of one chapter (lecture) compared with the other is the process to find the maximal similarity of one chapter (lecture) compared to all the chapters (lectures) of the other course.

Algorithm 1 Structural Similarity Algorithm

Input: Two Courses in Tree Structures A and B

Output: Structural Similarities in Different Levels from A to B

- 1: Initialize the set of structural similarities S as empty
 - 2: Compute the sets of chapters and lectures sub tree structures U_A, U_B, L_A and L_B from A and B
 - 3: Compute $Similarity(A, B)$ and add it in S
 - 4: **for** all u_i in U_A **do**
 - 5: **for** all u_j in U_B **do**
 - 6: Compute $Similarity(u_i, u_j)$
 - 7: **end for**
 - 8: Compute $Max(Similarity(u_i, u_j))$ and add it in S
 - 9: **end for**
 - 10: **for** all l_i in L_A **do**
 - 11: **for** all l_j in L_B **do**
 - 12: Compute $Similarity(l_i, l_j)$
 - 13: **end for**
 - 14: Compute $Max(Similarity(l_i, l_j))$ and add it in S
 - 15: **end for**
 - 16: Output S
-

Changes of usage $CH(UG_{S_1, l_1}, UG_{S_2, l_2})$ from l_1 to l_2 is one five-items set including the *increasing* or *decreasing* on NA , NP , ND , NO and US . Changes of *Usage score* reveal the *decreasing* or *increasing* of the general usage in the form of ranking, but does not reflect the *changes* on different parameters. The same changes on *usage score* may be due to different changes of four variables.

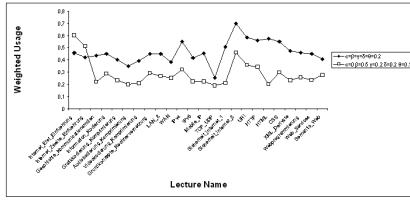


Figure 1. Usage Score on WWW lectures

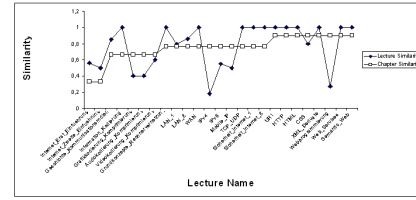


Figure 2. WWW course changes on lecture and chapter level

3 Mining Results Analysis on tele-TASK

We implement our mining methods on our web-based learning environment: tele-TASK, and the learning data includes: access logs from HTTP server and access logs from Helix server. Both kinds of access logs are taken from one semester 01.04.2006~31.07.2006 (LOG_I) and another semester 01.04.2005~31.07.2005 (LOG_{II}). Besides, we gathered the compositions of courses *WWW* for these two semesters. *WWW* from 2006 includes 26 lectures, while *WWW* from 2005 includes 31 lectures.

3.1 Data Preparation

We found that 80%~90% HTTP requests are made from crawlers or robots, while not from human. HTTP server logs related with web usage mining was widely discussed in [2,11]. In [12], we explained the work in details on data preparation in e-learning, which includes the heterogeneous learning usage data. Following table shows examples of reconstructed student browsing profiles.

Table 2. Example of Student Browsing Profile

Session	Learning Object	Type	Start Time	Duration	Operation
...
547	www12	0	31/May/2006:08:48:53	00:00:20	1
548	www12	1	19/Jun/2006:19:27:55	00:56:33	2
548	TL08	1	19/Jun/2006:20:31:55	00:21:42	6
...

3.2 Usage Information on Learning Lectures

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*. From LOG_{II} , we extracted out 948 learning sessions on *WWW*. To compute the average access duration of students on lectures, the various length of lectures have to be considered. Spending 25 minutes on a only 30 minutes long lecture shows more learning interest than costing 35 minutes on

a 100 minutes long lecture. In our experiments, we replaced *average access duration* with $(Average\ Access\ Duration)/(Length\ of\ Lecture)$. The distribution of usage scores on different lectures are shown in Figure 1. We show here two kinds of assignments on coefficients: $\{\alpha = \beta = \gamma = \delta = \theta = 0.2\}$ and $\{\alpha=0, \beta=0.5, \gamma=0.2, \delta=0.2, \theta=0.1\}$. The former overlooks the difference of different usage information on deciding the learning interest. The number of accessing live streaming version is much less important than that of accessing the edited version due to the short active period of live streaming. And further, the distributions of different usages help to decide the values of coefficients.

We also used the methods for mining frequent item sets to find if there exist some relations between the lectures during the same learning sessions [11]. We find that the first two lectures were always viewed together in all courses. 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.

3.3 Results of Similarity Comparison of Courses

Figure 2 shows the similarity comparison between the contents and organizations of two same titled lectures "WWW" from summer semesters 2005 and 2006. From this figure, it shows clearly that the similarity on the course level is not linearly decided by those on chapter or lecture levels, and also the similarity on the chapter can not linearly decided by those on lecture level. This further proves the necessity to compute the changes of learning interests on different levels. On the course level, the similarity of course *WWW* between 2005 and 2006 is 0.75. It is interesting that the similarity of first chapter between two years is 0.33, which is much lower than those of the two lectures in this chapter: $0.33 < \min(0.5, 0.56)$. This is because there was one more lecture in the first chapter in 2006 than in 2005. The similarity of chapter 3 is 0.9, but this hides the great changes in the lecture "Web Programming", of which the similarity is only 0.27.

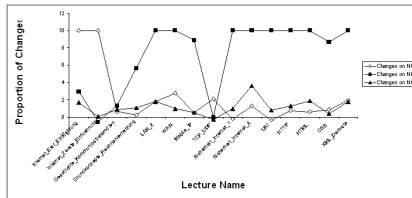


Figure 3. Changes of Usage on WWW lectures

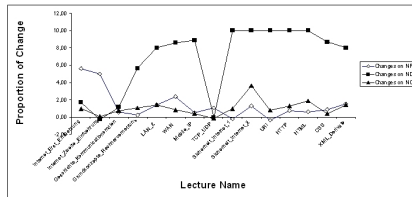


Figure 4. Changes of Weighted Usage on WWW lectures

The changes of usage on WWW from 2005 to 2006 is shown in Figure 3. Due to the popular acceptance by students on e-learning and efficient arrangement of teachers on tele-teaching 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 changes of learning interest integrating both changes on usage and lectures are shown Figure 4. We can see the difference between Figures 3 and 4. Though explosively increasing of learning interest 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.

From our investigations on web learning, we draw that the students have already been familiar to the basic knowledge such as "HTML" and "URI" before they choose this course from 2005 to 2006. We can adjust this course in the future: delete or compress the lectures on "HTTP", "HTML" and "URI", while enlarge the lectures about "Web Services", "Semantic Web" or other knowledge on the frontier of WWW.

4 Conclusion and Future Work

In this paper, we raised a method integrating the changes on usage and learning objects to measure the changes on learning interest on different levels. Currently we are con-

centrating on finding the right coefficients from statistic over huge learning data set and integrating the learning interest with the assignments of exercises and examinations.

References

- [1] S. Baron, M. Spiliopoulou. Monitoring the Evolution of Web Usage Patterns. Web Mining: From Web to Semantic Web, First European Web Mining Forum, 2003.
- [2] B. Berendt, M. Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. The VLDB Journal, 9: 56-75, 2000.
- [3] S. Chawathe. Comparing Hierarchical Data in Extended Memory. In Proceedings of the Int. Conf. VLDB, 1999.
- [4] J. Hardy, M. Antonioletti and S. Bates. e-Learner Tracking: Tools for Discovering Learner Behaviour. In Proceedings of Int. Conf. IASTED Web-Based Education, 2004.
- [5] R. McAlesse. Navigation and browsing in hypertext. In Hypertext: theory into practice, 1999.
- [6] A. Merceron and K. Yacef. TADA-Ed for Education Data Mining. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 7(1), 2005.
- [7] 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. AAI, 2005.
- [8] O. Nasraoui, C. Cardona and C. Rojas. Mining Evolving Web Clickstreams with Explicit Retrieval Similarity Measures. In Proceedings of Int. Conf. ACM WWW, 2004.
- [9] V. Schillings and C. Meinel. tele-TASK - Teleteaching Anywhere Solution Kit. In Proceedings of Int. Conf. ACM SIGUCCS, 2002.
- [10] C. Zinn and O. Scheuer. Getting to Know your Student in Distance Learning Contexts. In Proceedings of European Conf. TEL, 2006.
- [11] L. Wang and C. Meinel. Mining the Students' Learning Interest in Browsing Web-Streaming Lectures. In Proceedings of IEEE CIDM, 2007.
- [12] L. Wang and C. Meinel. Discovering Characteristic Individual Accessing Behaviours in Web Environment. In Proceedings of RSFDGrC, 2005.