Automatic Categorization of Lecture Videos: Using Statistical Log File Analysis To Enhance Tele-Teaching Metadata

Franka Grünewald*, Maria Siebert, Alexander Schulze*, Christoph Meinel*

Hasso-Plattner-Institute*, Institute of Computer Science
University of Potsdam
Potsdam, Germany
franka.gruenewald@hpi.uni-potsdam.de
maria.siebert@uni-potsdam.de
alexander.schulze@student.hpi.uni-potsdam.de
christoph.meinel@hpi.uni-potsdam.de

Abstract: Parsing user access log files for retrieving additional information is a well known approach to obtaining additional knowledge of a web site. Most research interests focus on identifying users and tracking user behaviour. In contrast, this paper concentrates on the statistical evaluation of all available log data. Therefore, special items of a web page are detected, categorized and the access data of these items is analysed. This paper shows the related process using the example of an e-learning web portal. It starts with preparing the log data and analysing it manually. Afterwards the data is categorized according to the findings of the analysis and the results proven by manually selected test data. Based on categories, it is possible to recommend tele-teaching objects with similar access data and classify them automatically to these categories.

1 Introduction

Online tele-teaching web portals have been in use for some years now. In addition to these online portals, recordings of lectures were created in order to facilitate time and place independent learning. With more and more content the distribution of the lectures becomes a problem. For that some universities use globally available systems like iTunesU or YouTube. Another approach is to set up a local portal for this service. This allows retaining control of the data. However, this leaves the burden of providing relevant metadata, which supports the user in finding the appropriate content, to the content providers. Because normally the number of users is not large enough it is not possible to just rely on user-generated metadata. Up to now this issue has been mostly handled by manually administrating the metadata.

With the amount of recorded lectures growing, this method will not be possible for all institutions anymore due to time and financial matters. This means that automatical methods for metadata harvesting have to be taken into account. Several approaches have been

published to get additional metadata that is generated by users. The problem is, that these local portals, that are mostly limited to e-learning content only, have a smaller user participation than the bigger, more general projects and the amount of user generated data is smaller. That is why approaches are researched to find other data sources. One idea is the usage of data logs. They provide data generated through visits from users. Parsing user logs is an old approach, which has been discussed in several ways before.

Looking at our user logs, we found out, that a large number of users only access a few pages. Most of them access the portal using search engines or direct links. So this access data is not usable for user tracking. Instead we looked at the accesses to different single pages without considering the pages visited before or after. This allowed us to use all data.

This paper will discuss one approach for detecting properties of different tele-teaching objects by comparing the user access data for their pages. Therefore, after giving a short description about the underlying project and the discussion of related work, we will elaborate on what we intend to do with the newly gained data, show how we prepared the data, describe the statistics we generated and give some examples for the properties extracted from that data. At the end, we will look ahead to future work.

2 About the Project

The web portal, which is the basis for our research, presents lectures recorded by a video lecture recording system. The system allows recording two video and one audio stream simultaneously. It has been in use since 2001. The result is more than 8000 lectures. Most of the lectures and conferences are about topics in computer science. These lectures are held by more than 1000 speakers and were organized in more than 250 series.

The latest version of the web portal, which is implemented in Django - a python web framework - is about 30 months old. During the relaunch of the web portal the whole structure changed. That is why we concentrate on these last 30 months in this work. The older data is not comparable with the new data because the overall structure of the data in the web portal has changed completely.

The portal reaches about 2000 users a week with many external users. Most of the students at the institute use the portal for learning, but a large number of users enter it using search engines or links on other web pages.

3 Related Work

There exist a lot of different publications about parsing user web access log files. The paper [Sud10] gives an overall description how log files can be mined. Furthermore, the users are identified for better analysis and an interface for data access is provided. The problem of finding sessions and user tracking are also discussed in [BTCF09]. This work furthermore elaborates how to detect patterns in the usage of web sites.

Other papers discuss this topic for special domains. In [WNLL03] a web log miner for product development is described. Even if this is distant from e-learning at first glance, the paper argues the need for research about the evaluation of log files in special domains.

The paper [DBG+07] describes the idea of using the user navigation patterns for generating topic directories. Therefore, the analysis of the user access data should help finding topic categories for the content.

In [WM07a] and [WM07b] the research area for analysing user access data is e-learning. The access path chosen by the user is analysed. The distance of different paths is measured. The access time on the learning videos is considered and two lectures series with an analogue curriculum are compared to detect differences in the learner's interest over time.

Another aspect of statistically analysing web documents is shown in [VD01]. This work does not use the web access data, but the data provided by the pages itself. Therefore, they defined their own metrics to find connections between different web pages.

Classification of web pages is an important part of information management and retrieval. The paper [QD09] describes how different web sites can be classified. Resulting improvements are shown, especially enhancements of the quality of search. The focus of this paper focuses on different web sites instead of the classification of pages of one site.

A use case for the extracted information of web log data is shown in [PCN08]. It describes, how the data can be used for a recommendation system. Recommendation systems provide great assistance in navigating big archives. Because we want to build up our own recommendation system, this work shows possibilities for the extension of the recommendation function. This aspect is be explained in more detail in the next section.

4 Usage of the categorization data

A similarity measure system is an important part of a recommendation system. There should be a combination of similarity measure results from different sources. An example where different similarity measures are combined is described in [SMM10]. It is possible to use the results gained by the analysis of the log files as a module for the overall calculation.

The following section will explain the categories defined for the automatical classification method described in this paper. For similarity measures, being in the same category creates a similarity for two objects. With a higher number of categories and associated entries for each category that similarity measure becomes more accurate. This approach also allows us to compare different results for similarity measures. This can help to improve our data and give valuable feedback for the automatic classification focused on in this paper.

The categorization can furthermore help to complete search tasks in a standard search interface or for serendipitous browsing. It can also be the basis for an automatic indexing of the content, which is desirable to speed up search and filtering tasks.

In order to find attributes that classify e-learning content to match a certain category, the data has to be prepared and analysed. Our methods to do so are described next.

5 Data Preparation and Categorization

We decided to use the Apache Log files for the statistical analysis, which allows us to use data of more than two years of user interaction. This section first explains how the log files are parsed and the data is prepared for analysis. Afterwards it is described how we define categories and manually cluster some example e-learning objects. Those two steps are based on experience with the subjects and their classification within the curriculum as well as the metadata gained with the recording and distribution of the e-lectures.

5.1 Parsing Log Files

The portal is set up on an Apache web server. Therefore, we have log files produced by the Apache server for every access of the server. Furthermore, we have special log files for https access to the web page. But https is not used for accessing the video archive, but only for user interaction commands. That is why we decided, that we do not have to parse the https files for our questions about user behaviour in the video archive.

On our web server we have administrated the following standard log file format.

```
1 IPADDRESS - COOKIE - [TIMESTAMP] "(GET|POST) URL PROTOCOLVERSION"
      RESPONSECODE DURATION "REFERRER" "USERAGENT"
```

Listing 1: Structure of Apache Log file line

Adding this data from the apache http log files to a database was the first step. Afterwards, the imported data had to be refined and prepared according to the following steps. The first is the Bot Detection. Bots are parsing a web page to gain data for search engines. They can be identified automatically using the user agent string or by their behaviour. They produce a large number of requests. Second the URLs need to be classified by type (e.g. image or webpage) and filtered for the files required. Third, internal data objects need to be detected, because only those can deliver more information for the analysis. Using Django functions it is possible to connect a URL with the represented data object.

Finally the database needs to be optimised. Optimisation needs to be done when the amount of data grows and the database requests become more complex. We have tables with more than 10 million entries therefore optimisation is required. Therefore, it is sometimes helpful to save redundant data, to avoid the usage of join clauses. This type of duplicate data allows us to get faster requests, but needs more checks for consistency.

5.2 Categories and their Attributes for the Categorization of E-Learning Objects

The tele-teaching content in the portal is organised in three different content types - series, lectures and segments, which can be understood as three different layers. The type of a content object can be determined by its URL. We divided the e-learning objects into their

different types. In this paper we focus on the type lecture series. A series is the collection of lectures. This can include the lectures of one subject held in one semester as well as all presentations of a conference. Each lecture in the system has one specific series it belongs to. Further content items can be analysed in the same way.

In this section we give a detailed description of the categories and their attributes we chose, based on the metadata available and the metadata that is considered interesting for filtering and search purposes (see sec. 4). This includes the type of content, how often it is recorded, the previous knowledge needed to take the course and the topic of the course. The goal is to find similar attributes in the samples that can be generalised. Those can then be used as basis for an automatic categorization algorithm.

(S) The *type of a series* is fixed even before the recording starts. The types we differentiate are events (e), conferences (c) - which also includes workshops and symposia as the number of content items does not allow any further refinement -, seminars (s) and lecture series (l), a normal course that runs a whole semester and ends with an examination. Because we utilize a university e-learning portal as sample project, a large part of the content is this type, because many lectures are recorded to provide better learning possibilities to the students.

(R) The usage of an e-learning object by the users depends on the *number of repetitions* of the series. If a series is recorded every year, older recordings may be looked at more seldom than a single recording which has only been used for many years. We differentiate in one time recording (o), sometimes rerecorded (s) and yearly repetition (y).

(K) The *knowledge needed for understanding a course* depends on the semester, the course is held. Normally, courses that are part of the bachelor in the first semesters are easier to understand than for example a master course. Therefore, it is necessary to differentiate between beginners (b), intermediate (i) and high previous knowlege (h) needed to be able to follow the course.

(T) The *topics* need to be separated, because most of our lectures are held in computer science and need to be distinguished from other topics. The three topic categories are: computer science (c), other specific topics (s) and other non specific topics (n).

In order to be able to manually analyse sample objects in the newly defined categories, we inspected different kinds of diagrams. These diagrams are presented in the next section.

5.3 Types of diagrams

We decided to create several diagram-types for each object to find typical repeating attributes for the categories. Each diagram gives another perspective on user access, although all are based on the same data.

5.3.1 Hits-over-time

The hits-over-time-diagram (see example in Fig. 1 top left) is the most typical diagram for presenting access-data. It presents the number of accesses in relation to the time. Therefore, the number of visits are grouped by weeks and the line displays the hits per week. The overall interval of data is visible beyond the line, which is important if a series was not available for the whole timespan of the evaluation.

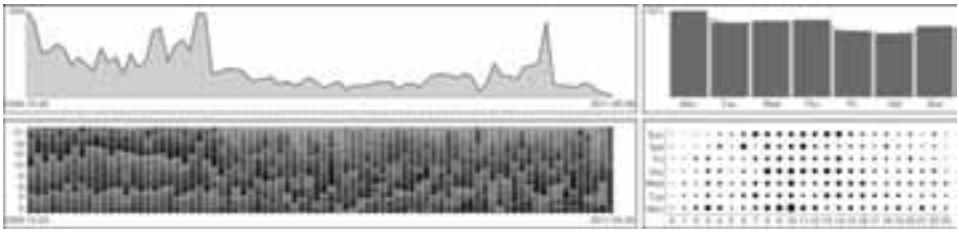

Figure 1: hits-over-time- (upper left), hits-within-a-week- (upper right), hours-over-time- (lower left) and hits-within-a-week-and-day (lower right) diagram for an object since its creation time

The object in figure 1 (in the upper left position) has most of its visits in the first semester after recording. You can also see some other peaks for the data, which could be special events.

5.3.2 Hits-within-a-week and Hits-over-the-day

The hits-within-a-week diagram (see fig. 1 top right) shows the number of visits for every single day in the week, especially the days with the most and the least visits. We summed up all visits for each weekday, which results in an absolute calculation. It is especially important to compare weekdays to weekends, because this could help finding out, if a series is watched during working times or in leisure time.

This diagram should be looked at together with the weekday-over-time diagram, which shows the changing access on weekdays over time.

The hits-over-the-day-diagram sets the number of hits in relation to the hours in a day. For example, it might be possible that lectures are often watched in the evening or at night, because this is the time, when students are often doing their homework and learning for tests. In contrast, conferences are most likely seen in the morning, because they are probably watched by employees.

5.3.3 Hits-within-a-week-and-day

The scatter plot diagram hits-within-a-week-and-day (see fig. 1 lower right grapics) shows the correlation between weekday and time of the day. It illustrates if a series is accessed at a special time on a special weekday, thus making it possible to find hotspots of viewing

interest. In this example, it is obvious that this series is watched often on Mondays at 10 o'clock but seldom on Sunday mornings.

Some lectures are viewed shortly after their recording. Therefore, we can find a hotspot of viewing for a special day and time. Other courses are viewed nearly constantly over time and weekdays, which generates a more diffuse image.

5.3.4 Other diagrams

We furthermore created weekdays-over-time-, hours-over-time- (see fig. 1 lower left image) and weekdays-over-hours-diagrams. These will not be discussed in further detail as the findings described in this paper are mostly based on the aforementioned diagram types. All data is visualized relatively in the diagrams shown, so that they are harder to interpret.

According to the attributes formulated in the last section different data samples are selected. This selection will be presented in the next section. Those are analysed manually with the help of available metadata. It is the goal to find similar attributes that are often re-occurring.

5.4 Sample Categorization

We manually selected 20 objects to categorize. When selecting these objects, we chose objects that have a high rate of views, mostly lecture series, and tried to select objects with a high coverage over all categories. In table 1 you can find the selected series objects including the results gained from manually analysing the lectures based on experiences of the evaluator and metadata of the lectures.

id	(S)	(R)	(K)	(T)	id	(S)	(R)	(K)	(T)	id	(S)	(R)	(K)	(T)	id	(S)	(R)	(K)	(T)
278	l	o	b	s	373	l	y	i	c	386	c	s	i	c	297	c	s	i	c
368	l	o	i	c	291	l	y	h	c	233	s	o	b	c	364	e	o	b	n
256	l	o	b	c	302	c	o	b	c	235	c	s	i	c	378	e	o	b	n
396	l	s	b	c	340	l	y	h	c	366	l	y	i	c	320	l	o	h	s
393	l	y	b	c	367	l	o	h	c	13185	s	o	b	c	13417	e	o	b	n

Table 1: Classification of series objects

The selection of seminars brings the problem, that the number of seminars in the portal is small and most of them were recorded in the last semesters. Therefore, achieving the goal to select different types of seminars from different periods of time was not possible. Also, this group is small so that we cannot be sure about the results for the seminars.

This first manual categorization of tele-teaching objects to the categories can now be used to deduce a more general assertion about attributes required to sort the e-learning objects into a specific category.

6 Analysis of the Attributes to the Categories

First, we will explain the categorization according to the series type and afterwards according to the knowledge needed. In the end we will explain the problems we had with the categories number of repetitions and topic.

6.1 Series Type

The clearest finding for differences in our set of diagrams were visible for the categorization of the type. From the diagrams for the different items, we could easily see, that the curve for the hits over time for lectures and conferences/events are obviously different.

In the portal the overall rule is that conferences and events are in fact viewed often shortly after the recording, but that the number of accesses decreases rapidly later on. On the other hand, lectures and seminars are watched regularly over a longer period of time, and even after the exams the course is watched regularly. Furthermore, we can see some exam peaks for lectures, which are missing for seminars. The example diagram in figure 1 (top left) shows a classic lecture. In the semester the course was held, there is a high number of accesses every week peaking around the time of the examination.

6.2 Knowledge needed

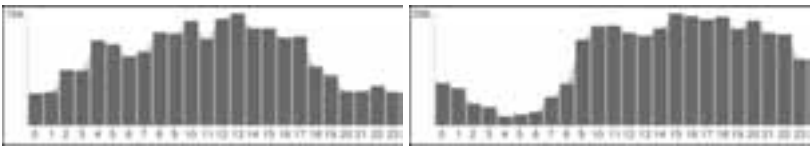

Figure 2: Diversification of visits throughout the day for a lecture that needs a lot of (left) and low (rigth) previous knowledge.

Another interesting fact is the diversification of visits over the day. Lectures which are particularly relevant for users with a lot of previous knowledge are watched in the morning hours more often (e.g. object 291 in fig. 2 on the left) than lectures which are particularly meant for beginners (e.g. object 396 in fig. 2 on the right). One explanation could be that beginner students often enjoy their student life and sleep long while advanced learners have to complete more tasks in a day, for example working, and therefore utilize different time slots.

To prove this hypothesis, we decided to define the hours between 3:00 and 5:00 as the hours in the morning, and the hours between 20:00 and 22:00 as the hours in the evening. Furthermore, we defined $n(h)$ as the number of views in the hour h. Afterwards, we calculated a skill-ratio sr for every test-object as follows.

$$sr = \frac{n(3) + n(4) + n(5)}{n(20) + n(21) + n(22)}$$

In table 2 we listed the test objects with the calculated skill-ratio. The events and conferences were filtered because they are not used by students for learning and influence the test results unnecessarily.

id	367	291	373	366	368	340	320	256	278	233	13185	393	396
(S)	l	l	l	l	l	l	l	l	l	s	s	l	l
(SR)	2.55	2.03	1.37	1.12	1.04	0.99	0.8	0.56	0.45	0.29	0.16	0.14	0.13

Table 2: Skill-Ratio (SR) for the lecture and seminar type objects with knowledge-category; ordered by Skill-Ratio.

We can use the skill-ratio as a strong factor for calculating the level of knowledge required. If $sr(s)$ is the skill-ratio of object s, we can classify objects with a skill-ratio $0 \leq sr(s) \leq 0.7$ as courses for beginners, with $0.7 \leq sr(s) \leq 1.7$ as courses for intermediate users, and with $1.7 \leq sr(s)$ as a series that requires more previous knowledge.

There are some objects which do not fit in the categories we have classified. This can be implicated by our own incorrect categorization (e.g. we had a hard discussion about the categorization of the object #320). Another reason can be the number of views or the exception that breaks the rule.

This factor seems to be good advice for the amount of knowledge required for a course if it is a lecture. Because a lot of data is needed for significant results (at least 1000 accesses in a period of at least 20 weeks), only the top viewed courses can be categorized this way.

6.3 Number of Repetitions and Topic

We had the problem that our data evaluation time was quite short. That is why we did not find many visible effects on repetitions and could not clearly define attributes for the number of repetitions. The only fact we could see is that if a lecture is not recorded a second time, the number of accesses to the old recording is still high, while with newer recordings it decreases. But this effect of the decrease could not be separated from other factors contributing to decrease, such as the course not being offered again.

In the diagram in fig. 1 (top left), the second exam peak is nearly as high as the first one, which indicates, that this course was not recorded again. This is the case, because normally a new recording flattens the second exam peak.

With the topic category, we had the problem that 90% of the content in our web portal is classified as computer science. Therefore, it is hard to find a series with another topic. And most of those have a low access rate and could not be used for evaluation, as explained further in the next section.

7 Evaluation

For the evaluation we selected overall 20 items we had not categorized so far. The selection was done for each block by the person who did not do the test interpretation. The 10 objects for each test person were anonymised and analysed by the criteria described beforehand. The diagrams, that have been explained in section 5.3, were used to manually analyse attributes of the categories based on the frequency distribution of the user access data.

A problem was the small number of usable data. Only about 80 series had enough accesses.

	categoritzation				test result			
#Test	(S)	(R)	(K)	(T)	(S)	(R)	(K)	(T)
1	l	o	i	c	l	o	b	c
2	e	o	b	c	l	o	b	c
3	l	s	b	c	l	s	b	c
4	l	s	i	c	l	y	i	c
5	l	y	h	c	l	y	h	c
6	e	o	b	n	l	o	i	c
7	l	y	i	c	l	y	i	c
8	e	o	b	n	c	o	b	n
9	c	o	b	c	c	o	i	c
10	s	o	i	c	c	o	b	c

	categoritzation				test result			
#Test	(S)	(R)	(K)	(T)	(S)	(R)	(K)	(T)
1	l	y	b	c	l	o	b	c
2	l	o	i	c	l	o	i	c
3	e	o	b	c	s	o	b	c
4	s	o	i	c	c	o	b	c
5	e	o	b	c	e	o	b	c
6	e	o	i	c	c	o	i	c
7	l	s	b	n	s	o	b	c
8	l	s	b	c	s	o	b	c
9	e	o	b	n	e	o	b	n
10	l	o	i	c	l	o	b	c

Table 3: First (left) and Second (right) evaluation for Classification of series objects

Test person one reports, that some of the results were difficult to decide. Because there is no good strategy for the topic, it is most times just agreed to be computer science as the most supposable value. Test #8 was chosen with another value, because of the late first peak, which happens normally, when non-computer science persons access a page.

With test objects #2 and #6 the problem was the low rate of overall access over the test period. This creates diagrams with only a low rate of access per week making a decision for the series type considerably difficult. Therefore, we could see that the overall number of access must be high enough to get good results.

The second test person had problems deciding on the type. It was hard to decide between conference and event as well as between lecture and seminar. Most lectures, which are not well categorised are old lectures, where the last exams phase was before our available start or are new lectures which do not show the next exam peak. The particular reason to classify a lecture as a lecture was the existence of the exam peak for the second test person.

Both tests show a high similarity for the category of knowledge. When there is a wrong answer, in most cases, the category of the knowledge level was hard to give and tended to the other value as well.

For the topic, the strategy of choosing computer science creates a high correctness of values, but shows the problem of low diversity in this category. That is why we think the results for this category are not usable.

In the whole statistical analysis we realized that it is possible to categorize the e-learning objects by user access data. The results are exact, which means that it is possible to create algorithms for the automatic categorization of e-learning objects. But of course it also shows that statistical user access analysis can produce some false data.

8 Conclusion

In this paper, we showed that it is possible to find properties of e-learning items and automatically categorize them according to these properties. Those properties could be detected using simple statistical analysis of the access pattern of users. The approach we used was the manual analysis of the frequency distribution of the data visualized in different types of diagrams. The most interesting result was that by looking at access time we were able to learn the previous knowledge required for a course. Furthermore, we are able to get the type of series by looking at the access data over time.

Those two results could be proven by evaluation. The number of repetitions and topic in the two categories could not be clearly defined with the data available. More diverse lectures from different fields and a longer period of test data would be needed to be able to state significant results in those two categories.

Since it is possible to automatically classify tele-teaching objects, those algorithms can help reduce manual administration. But more research has to be done with other categories, more evaluation needs to be done and more data used.

9 Future work

In order to verify our results, we will recheck our data following the current semester. We are especially interested if the influence of the required knowledge will still be visible or if it was an effect during a special period. Furthermore, we plan to carry out the same statistical analysis on other objects of our website. Also the video logs can be analysed for more information on the duration of the user visits for a special lecture. Therefore, we have to analyse the more complex streaming server logs. With knowledge about the lectures in each series, we can probably extract more information about the series as well.

Also important for improvement of the results is more activity in the portal. More possibilities for the users to interact with the portal need to be offered in order to get more data that we can analyse. For example, it should be possible for users to influence the knowledge value. If an object is rated as one that needs a lot of previous knowledge, the users should have the possibility to change this value by clicking on an *It's easy for me*-Button.

As proposed in section 4, we also want to use the data we gained from the analysis for building up a part of a similarity measure system. Therefore, we have to implement our categorization algorithm and store its results in the web portal database to evaluate this data when searching for content similarity.

References

[BTCF09] Murat Ali Bayir, Ismail Hakki Toroslu, Ahmet Cosar, and Guven Fidan. Smart Miner : A New Framework for Mining Large Scale Web Usage Data â Previous Heuristics and Related Work. In *Proceedings of the 18th international conference on World wide web*, pages 161–170, 2009.

[DBG+07] Theodore Dalamagas, Panagiotis Bouros, Theodore Galanis, Magdalini Eirinaki, and Timos Sellis. Mining user navigation patterns for personalizing topic directories. *Proceedings of the 9th annual ACM international workshop on Web information and data management - WIDM '07*, page 81, 2007.

[PCN08] Xueping Peng, Yujuan Cao, and Zhendong Niu. Mining Web Access Log for the Personalization Recommendation. *2008 International Conference on MultiMedia and Information Technology*, pages 172–175, December 2008.

[QD09] Xiaoguang Qi and Brian D. Davison. Web page classification. *ACM Computing Surveys*, 41(2):1–31, February 2009.

[SMM10] Maria Siebert, Franka Moritz, and Christoph Meinel. Distributed Recognition of Content Similarity in a Tele-Teaching Portal. In *2nd International Conference on Information and Multimedia Technology (ICIMT 2010)*, Hong-Kong, 2010.

[Sud10] G. Sudhamathy. Mining Web Logs - An Automated Approach. In *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India*, 2010.

[VD01] P. Vittorini and P. Di Felice. Statistical analysis of Web documents: a proposal and a case study. *12th International Workshop on Database and Expert Systems Applications*, pages 275–281, 2001.

[WM07a] Long Wang and Christoph Meinel. Detecting the Changes ofWeb Students' Learning Interest. pages 816–819, 2007.

[WM07b] Long Wang and Christoph Meinel. Mining the Students' Learning Interest in Browsing Web-Streaming Lectures. In *IEEE CIDM'07*, pages 194–201, 2007.

[WNLL03] Yew-Kwong Woon, Wee-Keong Ng, Xiang Li, and Wen-Feng Lu. Efficient Web log mining for product development. *Proceedings. 2003 International Conference on Cyberworlds*, pages 294–301, 2003.