XML Based Pre-processing and Analysis of Log Data in Adaptive E-Learning System: An Algorithmic Approach

Sucheta V. Kolekar, Radhika M. Pai(✉), and M.M. Manohara Pai

Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal University, Manipal, KA, India
{sucheta.kolekar,radhika.pai,mmm.pai}@manipal.edu

Abstract. E-learning has become the most popular way of delivering education and learning. Adaptive E-learning systems are the systems that adapt according to the requirements of the user. These systems should be capable of capturing the user preferences in terms of their learning styles and adapt the user interface accordingly. Web log analysis of the usage data can provide useful information regarding the learning styles. This analysis is extremely useful to develop an adaptive environment for the learner and at the same time for instructors to see how often their course contents are being used. In this paper a modified literature based approach is proposed where the learner's behavior is tracked by capturing the interactions with e-learning portal. The captured behavior will be stored in the form of sessions which will be grouped together to generate the sequence files in the XML formats. The learning styles have been identified by an algorithmic approach based on the frequency and time that the learners spend on various learning components on the portal. The approach is useful to provide an adaptive user interface which includes adaptive contents and recommendations in learning environment to improve the efficiency of e-learning. The learning style model used is Felder-Silverman Learning Style Model (FSLSM) to fit the learning styles into an adaptive environment.

Keywords: Adaptive E-learning · Data pre-processing · Usage patterns · XML · Felder-Silverman Learning Style Model

1 Introduction

Adaptive E-learning system deals with appropriate personalization and adaptation techniques in order to maximize the effectiveness of learning. It should be capable to detect or identify the user preferences and finally adapt it into the system. User preferences can be mapped to learning styles of the learner's. The Felder-Silverman Learning Style Model (FSLSM) is the popular learning style model available which emphasizes on various categories of learners. The adaptive user interface and contents should satisfy the requirements of FSLSM learners.

G. Vincenti et al. (Eds.): eLEOT 2015, LNICST 160, pp. 135–143, 2016.
DOI: 10.1007/978-3-319-28883-3_17

For adaptation purpose, the system should be designed to capture the usage data logs and identify the usage patterns. Web Usage Mining (WUM) applications are based on data which can be collected by capturing the usage data into log files. So in narrow sense WUM is also called as Web Log Analysis [1].

As the e-learning systems are deployed on web server, log data can be derived from several data sources such as server log files and databases. Web log files contain information about a learner's activity. Most of the times the activities captured in the log file produces non relevant information which may lead to wrong analysis of usage patterns. Web Logs Analysis phase of WUM is the important process to understand learner's way of using the e-learning system and based on the utilization usage can be made adaptive. The usage patterns are useful to understand learner's learning styles and it also can provide more information to instructors about the learner's behavior, and can give recommendations to learners by generating adaptive user interface components. To generate valid usage patterns data pre-processing becomes a necessary component which involves time to execute. In this paper, an algorithmic approach for pre-processing of captured log data is proposed. The log data should be pre-processed as per FSLSM sub categories so that the pre-processed data can be directly used for clustering the common user profiles. FSLSM is useful to define the various categories of learners in order to provide adaptation on the portal that satisfies each learner [2]. The data pre-processing phase is time consuming specially when e-learning system should include adaptation in real time. The pre-processed data should try to identify the learning preferences and useful to identify the categories of learners in order to provide adaptive user interface along with the contents.

2 Related Works

Zailani Abdullah et al. [3] proposed a sequential preprocessing model (SPM) and sequential preprocessing tool (SPT) as an attempt to generate the sequential dataset. The result shows that SPT can be used in generating the sequential dataset. They evaluated the performance of the developed model against the log activities captured from e-Learning System called myLearn. Cristbal Romero et al. [4] have done a survey on different types of educational environments and the data. They also discussed the main tasks and issues in the pre-processing of educational data, mainly using moodle. Navin Kumar Tyagi et al. [5] have carried out a survey about the data pre-processing activities like data cleaning, data reduction and related algorithms.

Shivkumar Khosla and Varunakshi Bhojane [6] have mentioned different web log files and pre-processing techniques. They introduced the concept of capturing different web log file while accessing the e-learning portal. Thanakorn Pamutha et al. [7] have focused on the preprocessing of the web log file methods that can be used for the task of session identification from web log file. The work in this study also produces statistical information of user session, such as: (1) total unique IPs; (2) total unique pages; (3) total sessions; (4) Session length and (5) the frequency if visited pages.

Felix Mdritscher et al. [8] have reported about an analysis study in order to approach a strategy towards the realization of LA functionality in the Learn@WU platform. In order to learn about the dependencies between Learning Management System's usage patterns and learning results, they examined the influence of 14 usage variables and the final grades of the participants of three large blended learning courses. Angel Cobo Ortega et al. [9] provides a brief overview of applying educational data mining (EDM) to identify the behaviour of learners in virtual teaching environments. Authors have used a fuzzy clustering algorithm to identify groups of learners based on their social interaction in forums and the temporal evolution of this classification.

Michal Munka and Martin Drlka [10] have focused only on the processes involved in the data preparation stage of web usage mining. They specified the inevitable steps that are required for obtaining valid data from the stored logs of the web based educational system. They compared three datasets of different quality obtained from logs of the web-based educational system and pre-processed in different ways. Chakarida Nukoolkit et al. [11] have performed exploratory data analysis and data mining on an e-Learning web log. The analysis uncovers the e-Learning users' usage behavior in accessing the content. The analysis also discovers e-Learning media popularity and usage patterns, and helps the institution fine tune future courseware, from strategic changes to the fine-grain of lesson content improvement.

Martin Cpay et al. [12] have described the applicability of different types of resources and activity modules in the e-learning courses and the worthiness of their usage. The presented ideas are supported by the outcomes of the questionnaire research realized within the e-learning study as well as the usage analysis of particular e-course. Yannis Psaromiligkos et al. [13] have examined the requirements for data mining facilities in LMSs. They have also described a new approach supported by a tool for analyzing learners' behavior in LMSs.

Shaily Langhnoja et al. [14] have given detailed description of how preprocessing is done on web log file and after that it is sent to next stages of web usage mining. Authors have mentioned areas of preprocessing including data cleansing, session identification, user identification, etc. Renuka Mahajan et al. [15] have discussed about the study which is conducted on usability and effectiveness of the e-content by analyzing the web log. Authors have evaluated different features of e-content that can lead to better learning outcomes for the learners, by understanding their navigational behaviors, their interaction with system and their area of interest. Nawal Sael et al. [16] have defined new static variables according to the Moodle-SCORM content tree and authors have applied more statistics and visualization techniques. In addition, authors have presented multidimensional graphics in order to understand users' accesses.

3 Proposed Methodology

Data pre-processing consists of various steps which are tedious and time consuming to implement in real-time application. Also same steps can not be suitable in

e-learning application for analysis. The proposed methodology describes about the issues to be addressed, parameters to be considered and algorithms to generate the XML files. The main objective of the proposed work is to capture the access patterns of the learners in W3C extended log formats and in database. The W3C log files give the usage of different pages accessed as per the learners login and page visit sequence. The database log gives the usage of different files accessed as per the course contents and time spent on that page. The implemented e-learning portal is a combination of learning components called pages and learning contents called files. Each learning component of portal and different types of contents are considered as part of learning objects. To map learning objects onto FSLSM dimensions all learning objects need to be labeled depending on preferences of each learner. The characteristics of FSLSM learner is studied in detail from the various research papers. Mapping of Learning Objects on FSLSM are shown in Table 1:

Table 1. Learning Objects Mapping as per FSLSM

Active	Videos, PPTs, Demo, Exercise, Assignments, Forum, Announcements, References
Reflective	PDFs, PPTs, Videos, Announcements, References, Email
Sensing	Examples, PDFs, Videos, Practical Material, Forum
Intuitive	PDFs, PPTs, Videos, Forum, TopicList, References, Assignments, Advanced topics link
Visual	Images, Charts, Videos, References
Verbal	PDFs, Videos, Email, Announcements
Sequential	Exercise, References, Assignments, Sequential Links, Email
Global	Topic Lists, References, Exercise, Assignments, PPTs, Forum

3.1 Assumptions/Requirements for Pre-processing of Web Log Data:

The following are the parameters considered:

1. Idle time spent on specific file or page is 10 min. If any learner not doing any activity on portal for 10 min then the session will terminate automatically and learner will get log-out from portal as per time oriented heuristic.
2. One specific learner's all sessions will be considered upto 30 min to generate final XML logs.
3. Unique session ids are maintained in every log record of learner.
4. Log records are sorted as per session time and different sessions of one learner are combined to understand usage patterns.

3.2 Parameters to be Considered XML Generation Algorithms:

The following are the parameters considered:

1. Session: Sequence of pages and files accessed by a learner on a particular website during a specified period of time. One sequence includes number of sessions with total time of 30 mins.
2. Frequency: Number of times specific file and page accessed by learners all sessions.
3. Time Spent: Total time spent on file or page by learner's all sessions.
4. Page Sequence: Pages and files accessed in specific order by learner. Page sequence is useful to identify the learning path of each learner.

```
INPUT: A finite set of Learners L = L1, L2,....,LN,
Sessions S =S1,S2,.....,SQ,
PageURL P=P1, P2,.....,PR
and FileURL F=F1, F2,.....,FX
OUTPUT: XML File
  initialize Session_Time ← 0, SessionLog_ID ← 0,
  SessionPageLog_ID ← 0, SessionFileLog_ID ← 0,
  PCount ← 0, PLogTime ← 0,FCount ← 0,
  FLogTime ← 0, Flag ← 0
  for each Sessions S_j where j ← 1 to Q do
      compute Session_Time = Start_Session − End_Session
      if "Session_Time " > 29min then
          get Learner_ID, SessionLog_ID from SessionLog file
      end if
  end for
  for each Session S_j where j ← 1 to Q do
      for each PageURL P_k where k ← 1 to R do
          get SessionPageLog_ID from PageLog file
          if SessionLog_ID == SessionPageLog_ID then
              get Learner_ID, PageURL, Page_ID, LogTime from PageLog
              file
              set Flag ← 1
          end if
      end for
      if Flag == 1 then
          for each PageURL P_k where k ← 1 to R do
              for each Learner L_i where i ← 1 to N do
                  if "PageURL "is accessed then
                      set PCount = PCount + 1
                      set PLogTime = PLogTime + LogTime
                  end if
                  create XMLTag for
                  Session_ID, Learner_ID, PageURL, PLogTime, PCount
              end for
          end for
      end if
  end for
  for each Session S_j where j ← 1 to Q do
      for each FileURL F_d where d ← 1 to X do
          get SessionFileLog_ID from FileLog file
          if SessionLog_ID == SessionFileLog_ID then
              get Learner_ID, FileURL, File_ID, LogTime from FileLog file
              set Flag ← 2
          end if
      end for
      if Flag == 2 then
          for each FileURL F_d where d ← 1 to X do
              for each Learner L_i where i ← 1 to N do
                  if "FileURL "is accessed then
                      set FCount = FCount + 1
                      set FLogTime = FLogTime + LogTime
                  end if
                  create XMLTag for
                  Session_ID, Learner_ID, FileURL, FLogTime, FCount
              end for
          end for
      end if
  end for
```

Algorithm 1. XML File 1 Generation Algorithm

```
INPUT: A finite set of Learners L = L1, L2,....,LN,
Sessions S =S1,S2,.....,SQ,
PageURL P=P1, P2,.....,PR
and FileURL F=F1, F2,.....,FX
OUTPUT: XML File

  initialize Session_Time ← 0, SessionLog_ID ← 0,
  SessionPageLog_ID ← 0, SessionFileLog_ID ← 0, IsPage ← 0
  for each Sessions S_j where j ← 1 to Q do
      compute Session_Time = Start_Session − End_Session
      if "Session_Time " > 29min then
          get Learner_ID, SessionLog_ID from SessionLog file
      end if
  end for
  for each Session S_j where j ← 1 to Q do
      for each PageURL P_k where k ← 1 to R do
          get SessionPageLog_ID from PageLog file
          if SessionLog_ID == SessionPageLog_ID then
              get Learner_ID, PageURL, Page_ID, LogTime from PageLog
              file
              set IsPage ← TRUE
          end if
      end for
      if IsPage ==' TRUE' then
          for each PageURL P_k where k ← 1 to R do
              for each Learner L_i where i ← 1 to N do
                  if "PageURL "is accessed then
                      create XMLTag for Session_ID, Learner_ID, PageURL
                      set Session_ID, Learner_ID, PageURL in
                      "UserActivities "table
                  end if
              end for
          end for
      end if
  end for
  for each Session S_j where j ← 1 to Q do
      for each FileURL F_d where d ← 1 to X do
          get SessionFileLog_ID from FileLog file
          if SessionLog_ID == SessionFileLog_ID then
              get Learner_ID, FileURL, File_ID, LogTime from FileLog file
              set IsPage ←' FALSE'
          end if
      end for
      if IsPage ==' FALSE then
          for each FileURL F_d where d ← 1 to X do
              for each Learner L_i where i ← 1 to N do
                  if "FileURL "is accessed then
                      create XMLTag for Session_ID, Learner_ID, FileURL
                      set Session_ID, Learner_ID, FileURL in
                      "UserActivities "table
                  end if
              end for
          end for
      end if
  end for
  for each Learner L_i where i ← 1 to N do
      set UserActivities Orderby Date and Time
  end for
```

Algorithm 2. XML File 2 Generation Algorithm

The log data has been generated in standard XML format based on Assumptions mentioned above. The XML file can directly be useful as input for clustering algorithms. Two different XML files are generated: File1 contains User session data of pages/files, time spent on page/file in each session and frequency of accessing the pages/files. File2 contains page and file sequence of each learner called web sessions.

3.3 XML File1 and File2 Generation:

Session can be described as the time spent on portal by a learner from the moment he/she logged in to the moment he/she logged out. In each session of each learner is combined and identified the time spent of each page/file separately. Session also describes how many time each learner accessed page/file. The total time spent on each page and file as well as frequency of accessing page and file is converted into XML tags with respect to unique learner id as shown in Algorithm 1. Session can also describe as sequence of pages and files accessed by each learner in specific order as shown in Algorithm 2.

4 Experimentation Details

The www.mitelearning.com portal is made available for second year engineering learners and some interested learners for the Android for Beginners online course. The topics are divided into three categories based on prerequisite concept, main concept and advanced concept. The contents for the topics are made available for the learners through the portal in different file formats such as text (doc/pdf), video (mpeg/mp4) and demo (ppt/pptx). The learners can also go through the exercise modules that has been provided for each topic. The learners can also make use of different learning components that has been provided as pages to explore more about a specific topic such as Announcement, Email, Assignments, References, Exercises, TopicList, MyAccount, Forum etc. Around 45 learners have registered in portal out of which 30 learners have accessed the portal for two months. The log details of learners who are accessing the portal are tracked and stored into W3C log files as well in the database of SQL server. In addition to that, the time spent and the files accessed by a specific learner in a particular session has been captured and stored in the database. The stored logs in database are classified into three types: Learners SessionLog, Learners PageLog and Learners FileLog.

5 Results and Discussion

5.1 Pre-processing of Captured Usage Data

The application which generates both XML files is implemented using Microsoft Visual Studio 2010 and Microsoft SQL server 2005. The application fetches session data of each learner from databases and IIS log files.

The resultant XML files can be saved on disk. The XML file generates different tags for each record which will be easy for understanding. The XML file 1 and file 2 is as shown in Figs. 1 and 2 respectively. In Fig. 2 XML file of each learner is having unique SessionId with SequenceId (it is sequence of activity number) and each learner can have multiple SessionId with SequenceId. The XML file 1 is generated to identify the time spent and frequency of accessing each file and page by learner in each session. The XML 2 file generates the sequences of accessing files and pages as per timestamp in combined sessions of

Fig. 1. XML file 1

Fig. 2. XML file 2

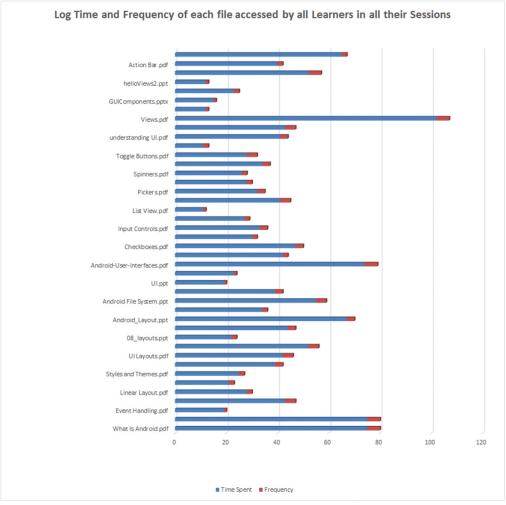

Fig. 3. Graph 1: Time spent and frequency of each file accessed by all learners

each learner. These both XML files are useful to identify common user profile based on aligned sequences, time spent and frequency (Figs. 1 and 2).

5.2 Analysis of Pre-processed Usage Data

After pre-processing the usage data and after generating XML files, usage data is analyzed to identify different statistical information. One of the important information is, how many times (Frequency) each learner accessed file and how much time spent(Log Time in Minutes) on each file by all learners in all their sessions. The graph 1 shows the information about more than 100 files.

6 Conclusion and Future Work

The use of personalized and adaptive e-learning system has become increasingly important in recent years, with extensive research being devoted to finding different ways of tailoring the learning styles for individual students. The work focuses on grouping the session details obtained from different log files. Different algorithms have been implemented to analyze the session log details of learners. The captured web log in the IIS and Database is an important source to identify the learning styles of learners. The learner's session has been considered as the total number of learning objects accessed by that specific learner. The captured database log data consists the details related to pages and files accessed by a learner as per unique session identifier allotted to the learner. The sessions are grouped together based on the time spent and frequency of the accessed learning objects. This helps to generate the sequences of learning objects. The FSLSM model is used to map the sequences into the learning styles. The log data has been converted into the standard XML format for clustering and learning path optimization.

References

1. Kolekar, S., Sanjeevi, S.S., Bormane, D.: Learning style recognition using artificial neural network for adaptive user interface in e-learning. In: Proceedings of IEEE Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–5. IEEE (2010)
2. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. Engr. Edu. **78**(7), 674–681 (2011)
3. Chiroma, H., Herawan, T., Deris, M.M., Abdullah, Z.: A sequential data preprocessing tool for data mining. In: Murgante, B., et al. (eds.) ICCSA 2014, Part III. LNCS, vol. 8581, pp. 734–746. Springer, Heidelberg (2014)
4. Romero, C., Romero, J.R., Ventura, S.: A survey on pre-processing educational data. In: Peña-Ayala, A. (ed.) Educational Data Mining. Studies in Computational Intelligence, vol. 524. Springer, cambridge (2013)
5. Tyagi, N.K., Solanki, A., Tyagi, S.: An algorithmic approach to data preprocessing in web usage mining. Int. J. Inf. Technol. Knowl. Manage. **2**(2), 279–283 (2010)

6. Khosla, M.S., Bhojane, M.V.: Capturing web log and performing preprocessing of the users accessing distance education system. Int. J. Mod. Eng. Res. (IJMER) **2**(5), 3128–3130 (2012)
7. Pamutha, T., Chimphlee, S., Kimpan, C., Sanguansat, P.: Data preprocessing on web server log files for mining users access patterns. Int. J. Res. Rev. Wirel. Commun. (IJRRWC) **2**(2), 92–98 (2012)
8. Mödritscher, F., Andergassen, M., Neumann, G.: Dependencies between e-learning usage patterns and learning results. In: Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, pp. 24:1–24:8. ACM (2013)
9. Ortega, A.C., Blanco, R.R., Diaz, Y.Á.: Educational data mining: user categorization in virtual learning environments. In: Espin, R., Pérez, R.B., Cobo, A., Marx, J., Valdés, A.R. (eds.) Soft Computing for Business Intelligence. Studies in Computational Intelligence, vol. 537, pp. 225–237. Springer, Heidelberg (2014)
10. Munk, M., Drlík, M.: Impact of different pre-processing tasks on effective identification of users behavioral patterns in web-based educational system. Procedia Comput. Sci. **4**, 1640–1649 (2011)
11. Drlik, M., Munk, M.: Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining. In: Cherifi, H., Zain, J.M., El-Qawasmeh, E. (eds.) DICTAP 2011, Part I. CCIS, vol. 166, pp. 60–74. Springer, Heidelberg (2011)
12. Nukoolkit, C., Chansripiboon, P., Sopitsirikul, S.: Improving university e-learning with exploratory data analysis and web log mining. In: 2011 6th International Conference on Comput. Sci. Edu. (ICCSE), pp. 176–179. IEEE (2011)
13. Cápay, M., Balogh, Z., Boledovičová, M., Mesárošová, M.: Interpretation of questionnaire survey results in comparison with usage analysis in e-learning system for healthcare. In: Cherifi, H., Zain, J.M., El-Qawasmeh, E. (eds.) DICTAP 2011 Part II. CCIS, vol. 167, pp. 504–516. Springer, Heidelberg (2011)
14. Psaromiligkos, Y., Orfanidou, M., Kytagias, C., Zafiri, E.: Mining log data for the analysis of learners behaviour in web-based learning management systems. Oper. Res. **11**(2), 187–200 (2011)
15. Langhnoja, S., Barot, M., Mehta, D.: Pre-processing: procedure on web log file for web usage mining. Int. J. Emerg. Technol. Adv. Eng. **2**(12), 419–423 (2012)
16. Mahajan, R., Sodhi, J., Mahajan, V.: Usage patterns discovery from a web log in an indian e-learning site: a case study. J. Edu. Inf. Technol., 1–26 (2014). doi:10.1007/s10639-014-9312-1