Web Usage Mining for Building an Adaptive e-Learning Site: A Case Study

Renuka Mahajan[1*], J. S. Sodhi[2], Vishal Mahajan[3]

[1] AIIT, Amity University, Noida, UP, India.
[2] AKC Data Systems, Delhi, India.
[3] HCL Technologies, Noida, UP, India.

*Corresponding author. email: rmahajan@amity.edu
Manuscript submitted July 10, 2014; accepted August 29, 2014.
doi: 10.7763/ijeeee.2014.V4.343

Abstract: Recent trends in adaptive e-learning, incorporate the use of data mining techniques, to deliver learning according to learners' requirements. In this paper, web usage mining is realized as a case study on an Indian e-learning site. The objective of this paper is the analysis of the web log data of an eLearning system and the deduction of useful conclusions. The web mining process consists of data collection, preprocessing, applying data mining techniques, followed by pattern analysis. It would help in identifying those interaction sequences which are most or least frequent and to explore how to reorganize placement of content and assessments on the basis of usage and learner assessment data, thereby making an e-learning site adaptive to deliver tailored e-content, as required by the learner.

Key words: Adaptation, personalization, sessionization, web usage mining.

1. Introduction

E-learning refers to the employment of information and communication technologies for development and delivery of learning [1]. E-learning has many synonyms viz. online education, technology-enhanced learning (TEL), multimedia learning, computer-based instruction (CBI), computer-based training (CBT), virtual learning environments (VLE), m-learning, and digital educational. These alternative names simply emphasize a particular component or a delivery method. Subsequently, the development of e-learning has rapidly progressed to include many alternative e-learning techniques, such as knowledge sharing or links to resources via social media sites, viewing online lectures etc.

While there are major benefits for all concerned, e-learning continues to pose challenges to providers to constantly develop new strategies for learning process. With the increased use of e-learning in schools and universities, there is a huge demand for better interpretation and implementation of e-learning platforms. According to [2], in e- learning, learners face many challenges such as lack of flexibility of the site, lack of adaptability towards learners' needs, lack of effective design of electronic content. This lack of adaptive learning environments or ***an*** environment with adaptive features is partly due to the concepts "one-size-fits-all". Hence the problem with e-learning sites is that it cannot teach learners in accordance to their aptitude and provide adaptive material.

E-learning systems generate huge amount of data. This information has the potential in understanding existing gaps to improve an e-learning site. Using this knowledge, e-learning environment can be made to

adapt to learners' needs to enhance their learning experience. This is why adaptation of e-learning systems to an individual or to a group based is the next step in the evolution of the e-learning systems.

Strategies for designing good adaptive e-learning site are one of the great challenges. For the purpose of this study, the focus will be on mining the web data containing the interactions of learner. "While a large number of organizations have adopted e-learning programs, far fewer have addressed the usability of their learning applications. As pointed out by [3] that "If organizations have to fully benefit from their investments, then more attention should be devoted, to ensure the usability of e-learning application". Data mining techniques can be used to extract knowledge in e-learning domain through the analysis of the data generated by their learners. The main objective becomes analyzing the patterns of system usage by teachers and learners [4]. Web mining [5] is the application of data mining techniques to extract knowledge from web log data. Oren Etzioni was amongst the first to define the term of Web mining in his paper [6].

In general, web-based educational systems have a lot of information recorded in log files, for example, interactions between students and the online learning systems, details of student successes and failures, student grades, and knowledge levels etc. In e-learning, there's much information available about not only learner's interaction rather about activities, such as reading, writing, taking assignments and communication with peers. Analyzing the server logs and the history list can help to understand the user behavior and the web structure, thereby improving the design of the website. It reveals interesting access patterns that can be used to restructure sites, determine effective advertising locations, and planning specific selling strategies.

- One can find information such as-
- Which components /web pages are most/least used?
- Which events are the most frequent?
- Providing real time navigation recommendations for e-learners
- Most and least frequently referenced.
- For educationists taking decisions on matters such as restructuring or reorganizing the site structure according to previous learners' usage.
- Creating access shortcuts for interested pages.

Lately, many researchers have worked to exploit the potential of data mining in e-learning applications but they don't show the detailed empirical results. The present study continues this work.

Web Mining: Web Mining [7] is of three types: - web content mining, which is the process of extracting useful information from the contents of web documents; web structure mining is the process of discovering structure information from the web; and web usage mining, which is the process of discovering meaningful patterns from data generated by client-server transactions stored in web logs. Hence, web usage mining is the process of finding out how the user uses the Internet.

Web data preprocessing is the first step in Web Mining. It is necessary to convert the data to an appropriate form (modified data) for solving a specific educational problem. This includes choosing what data to collect, focusing on the questions to be answered, and making sure the data align with the questions. It includes tasks such as data cleaning, user and session identification, page view identification etc. After data cleaning, filtering, pre-processing, integrating from multiple sources, we transform the integrated data into a relational database or a warehouse, suitable to be used as input to various web mining algorithms.

We organize this paper as follows: We describe our methodology in Section 2. We summarize the experimental results in Section 3, followed by discussions in Section 4. In Section 5, we provide our pointers for future work.

2. Methodology

We conducted experiments on a real world data set. Once the course is undertaken, the cumulative number

of web log data for each learner was downloaded and compiled for all learner clicks.

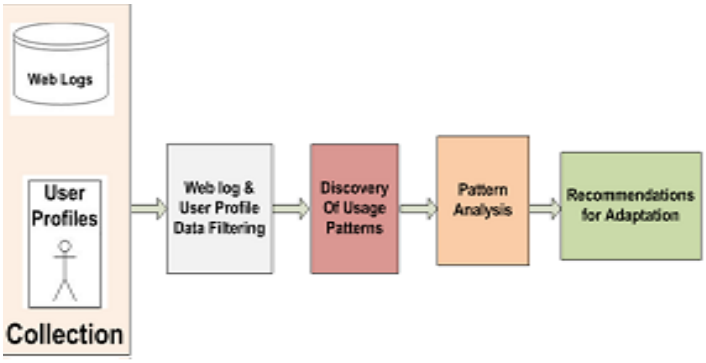

Fig. 1. Methodology for WUM for the e-learning site.

Population and Sampling used: This case study is based on an e-learning system designed for students of classes from 6th to 10th. It provides study planner, notes, reports to track progress, lesson summary for quick reference; previous years' board papers and practice tests on Biology lessons for CBSE, ICSE and 18 other Indian State Boards, to improve grades and clear learners' concepts. The sample of 3561 records was selected from a population of around 5000 records from the actual web usage log records from learn-next portal from Dec 1 2012 to Dec 31 2012.

Measurement instruments design and development: For the experiments, secondary data i.e. web log data was used. The dimensions consisting of usage and performance data were carefully selected from this web log data. The series of steps followed is shown in Fig. 1

2.1. Data Collection

In data collection, three sources of data were identified.

- The Server Side Data - The learners' traversal paths recorded in Web log at the web server.
- The Client side Data - The client side application or the client-side cookies.
- The proxy side Data- The data is gathered in a log file. It is useful to characterize a group of users using the same proxy server.

In this study, we specifically chose the server side collected data. A web log is a file to which the Web Server writes information, each time a user requests a resource from that particular site. The type of data in the web log consists of: IP address of the client, the software used by the client to navigate, time, date of request, IP address of the client, the resource requested, mode of request(POST/GET), the errors returned (if any) and the server-side click streams. The click stream is defined as an ordered sequence of pages visited by a user, in a session. This click stream data forms the basic element of Web Usage Mining.

The data obtained from various sources can be divided into the following groups.

1) Usage Data: The log data collected automatically by Server represents the navigational behavior of learners. Each hit, represents a single entry in the log. Each log entry consists of time, date of request, IP address of the client, the resource requested, mode of request(POST/GET), HTTP status ,no of hits, cookie etc. This data needs to be transformed and aggregated at different levels of abstraction. In WUM, the most basic level of abstraction is the page view. A page view is an aggregated representation of a collection of Web objects contributing to display on a user's browser, resulting from single user action (e.g. click through). At the user level, most basic unit of abstraction is session. A session is the sequence of page views by a single user during a single visit.
2) Content Data: It includes static HTML/XML pages, multimedia files, collection of records from operational databases.

3) Structure Data: Web designer's view of content organization, within the site e.g. HTML/XML pages can be represented as a tree structure forms the structure data.
4) User Data: It includes profile data and the demographic information about registered users, user ratings etc.

Some databases are insufficient, inconsistent and have noisy data. Hence pretreatment is essential to make this data consistent.

2.2. Data Preparation and Preprocessing

This was applied to retrieve this suitable data set from raw web log records, to which various data mining & statistical techniques could be applied. Web data preprocessing included tasks like data cleaning, user and session identification and page view identification. After data cleaning, filtering, pre-processing, integrating from multiple sources, we transformed the integrated data into a warehouse, suitable to be used as input to various data mining algorithms. In this data set, the entry in the session-page table associated with the specific page in a given session is determined by the number of web page hits. The set of tasks in usage data preparation phase, used to restore learners' activities in a reliable and consistent way are-

1) Data Cleaning: The main aim is to remove unnecessary entries in the web log. The irrelevant information usually is:
   - All image entries in the log having extensions .gif, .jpeg, .css etc.
   - The records having failed HTTP status code.
   - Robot-access patterns can be detected and removed.

   In the rest of the database, each row represents learner sessions.
2) Learner and Session Identification: The main purpose is to distinguish among different learners. In learner identification, one needs to identify who accessed web site and what pages were accessed. In session Identification, these page accesses of each learner are divided into individual learner sessions as shown in Fig. 2. Hence, a session is a series of web pages a learner browses in a single access. For this reason each row is processed in the way that it provides navigated web pages.
3) Path Completion: This is performed after sessionization. There could be many reasons for access not recorded such as local cache, proxy servers & browsers' 'back' button etc. Therefore, it is important to find out missing pages in the access path. From the combination of IP and Agent fields from the partly processed log file, we partition it into activity records of separate users.

Time	IP	URL	Ref	Agent
0.05	2.3.4.5	A	B	IE8;Win2k
0.08	1.2.3.4	B	--	IE8;WinXP;SP1
0.13	1.2.3.4	C	A	IE8;Win2k
0.14	1.2.3.4	B	C	IE8;Win2k
0.17	1.2.3.4	E	--	IE7;Win2k
0.16	2.3.4.5	C	B	IE8;Win2k
0.21	2.3.4.5	D	A	IE7;WinXP;SP1
0.21	2.3.4.5	A	--	IE8;WinXP;SP1
0.24	1.2.3.4	E	C	IE8;WinXP;SP1
0.24	1.2.3.4	C	A	IE8;Win2k
0.35	2.3.4.5	B	B	IE8;Win2k
0.59	1.2.3.4	D	C	IE7;WinXP;SP1
1.21	1.2.3.4	E	B	IE8;Win2k
1.25	1.2.3.4	A	D	IE7;WinXP;SP1
1.26	2.3.4.5	C	--	IE8;WinXP;SP1
1.27	2.3.4.5	F	C	IE7;WinXP;SP1
1.26	1.2.3.4	F	C	IE8;Win2k
1.31	1.2.3.4	B	A	IE8;Win2k
1.37	1.2.3.4	D	B	IE7;WinXP;SP1

	Time	IP	URL	Ref
	0.05	2.3.4.5	A	B
	0.08	1.2.3.4	B	--
	0.13	1.2.3.4	C	A
User 1	0.14	1.2.3.4	B	C
	0.17	1.2.3.4	E	--
	0.16	2.3.4.5	C	B
	0.21	2.3.4.5	D	A
	0.21	2.3.4.5	A	--

Fig. 2(a). User identification using browser type and IP address.

Session 1	0.24	1.2.3.4	E	C	IE8;WinXP;SP1
	0.24	1.2.3.4	C	A	IE8;Win2k
	0.35	2.3.4.5	B	B	IE8;Win2k
	0.59	1.2.3.4	D	C	IE7;WinXP;SP1
Session2	1.21	1.2.3.4	E	B	IE8;Win2k
	1.25	1.2.3.4	A	D	IE7;WinXP;SP1
	1.26	2.3.4.5	C	--	IE8;WinXP;SP1
	1.27	2.3.4.5	F	C	IE7;WinXP;SP1

Fig. 2(b). Example of identification of sessions.

Thus through data pre-processing, web log can be transformed into a dataset that can be easily mined. Thus the data was preprocessed and transformed. Pedagogical data as in [8] i.e. the page visited, the time of access, the referred page, the time spent viewing the page etc. were collected.

For our study, following fields as shown in Table 1were finally filtered - Total Attempts, Unique Attempts, total assessments - fail/pass, Assessments Attempted, Marks Scored, Average Time Taken (sec.) and the complete navigation path (Class, Subject, name of the Chapter, and the Lesson's Name) from 3561 web log sessions. This forms the final data base to be used as input for all the methods.

Table 1. Attribute Definitions for Each Learner

Name	Description
Average time	It is the average of the time spent by students on the specific topic, within a subject, in the selected data sample registered for the course.
Average Score	It is the average of the marks scored in assessment of each topic, within a subject, by the students in the selected data sample registered for the course
Total Attempts	These are the numbers of times a specific topic within a subject is referred by students in the selected data sample registered for the course
Marks Scored	These are the marks the student obtained in the assessment of a specific topic within a subject
n_assessment	Number of assessments done
n_assess_p	Number of assessments passed
n_assess_f	Number of assessments failed

2.3. Applying Data Mining

Finally, OLAP analysis & various statistical methods are applied on this web log data in SPSS 16. Statistical analysis are the common methods to give statistical description (trend and periodicity) of pattern based on frequency table, mean, median, standard deviation, histogram etc.

2.4. Pattern Analysis

The patterns thus obtained can be used to discover interesting information. It can be used for various purposes e.g. Web Site/Page Improvements, Additional Topic or Product Recommendations, Web Personalization, User Behaviour Studies etc. [9].

3. Results

Following patterns were identified after statistical analysis to produce useful information for decision-making for student of grade 10th for subject Biology. Following patterns gives learners' access overview.

3.1. Topics Frequently Accessed

3.3.1 Inference

The above log analysis in Fig. 3 clearly reflects the maximum use of following topics on the web site - Life Processes ,The Structure and Functional Unit of Life, Basic Biology ,Health And Hygiene , Life's Internal Secrets ,Cell - The Unit Of Life and Our Environment.

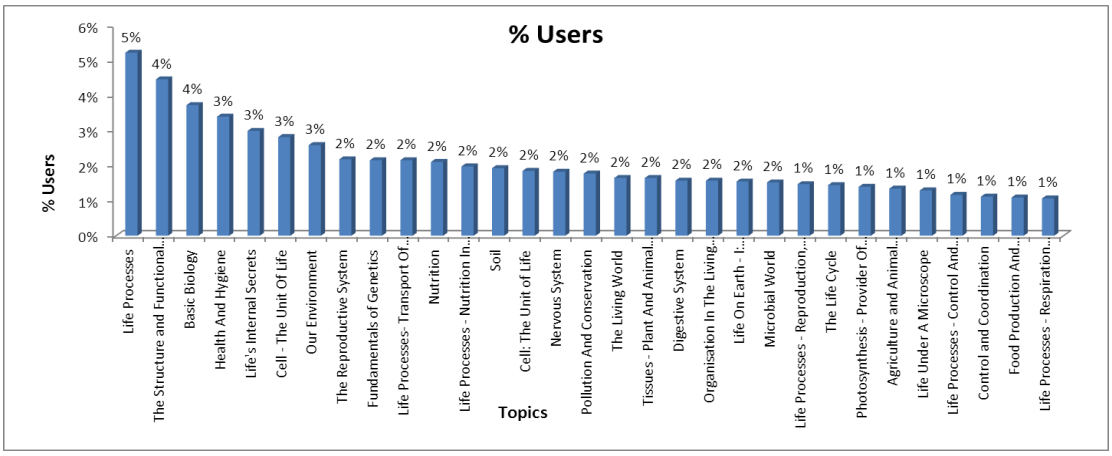

Fig. 3. Frequently accessed topics in biology.

However, least accessed resources are The Reproductive System ,Fundamentals of Genetics ,Life Processes- Transport Of Materials In Animals And Plants, Nutrition , Life Processes - Nutrition In Animals And Plants, Soil , Cell: The Unit of Life ,Nervous System , Pollution And Conservation ,The Living World ,Tissues - Plant And Animal Tissues, Digestive System , Organization In The Living Things – I, Life On Earth - I: Characteristics Of The Living Beings, Microbial World , Life Processes - Reproduction, Growth And Development, The Life Cycle , Photosynthesis - Provider Of Food For All Agriculture and Animal Husbandry, Life Under A Microscope ,Life Processes - Control And Coordination , Control and Coordination, Food Production And Management – I and Life Processes - Respiration And Excretion In Animals And Plants. It implies that there should be greater focus on above mentioned topics in order to increase learner access.

3.2. Average Time Spent on Each Topic

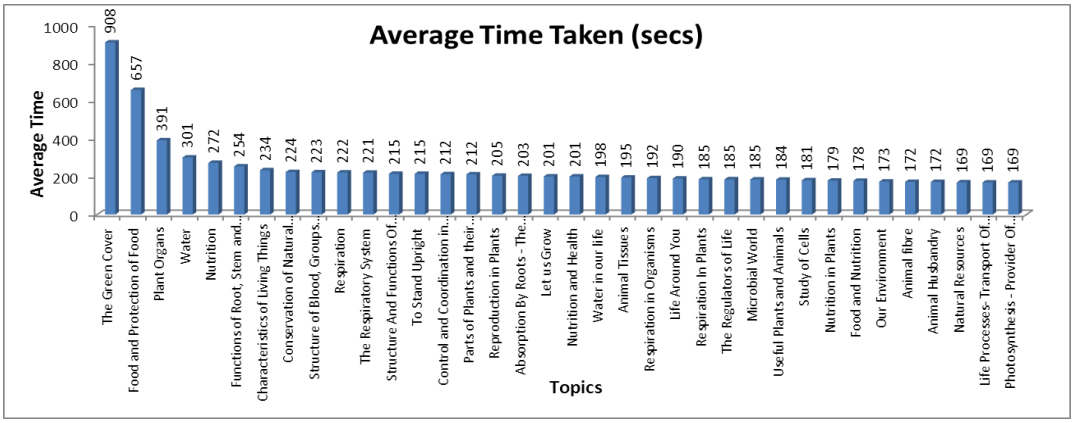

Fig. 4. Topic wise average time spent in biology.

3.2.1 Inference

As shown in Fig. 4, learners have spent considerable amount of time on topics - The Green Cover, Food & Protection of Food, Plant Organs and Water.

However, very less amount of time has been spent on topics- Nutrition ,Functions of Root, Stem and Leaf, Characteristics of Living Things, Conservation of Natural Resources, Structure of Blood, Groups and Blood

Transfusion, Respiration ,The Respiratory System ,Structure And Functions Of Organisms – II, To Stand Upright ,Control and Coordination in Organisms, Parts of Plants and their Structure, Reproduction in Plants , Absorption By Roots - The Processes Involved, Let us Grow ,Nutrition and Health , Water in our life , Animal Tissues ,Respiration in Organisms , Life Around You , Respiration In Plants , The Regulators of Life , Microbial World , Useful Plants and Animals , Study of Cells , Nutrition in Plants ,Food and Nutrition , Our Environment, Animal fiber , Animal Husbandry and Natural Resources. Studies have been done by [10] to explore in detail, the ways in which students engaged with a resource and to evaluate the ways and extent to which it might enhance their learning.

3.3. Average Assessment Score on Each Topic

3.3.1 Inference

Data representation in the Fig. 5 indicates that after accessing the site, the learners' have scored the highest percentile in topic – Nutrition in Plants, Food and Protection of Food, Cellular Organization etc.

However learner's had scored less on topics like Plant Tissues , Life Processes - Reproduction, Growth And Development , Animal Husbandry and Organization in the Living Bodies. Hence more focus should be given to these topics to improve learner's understanding.

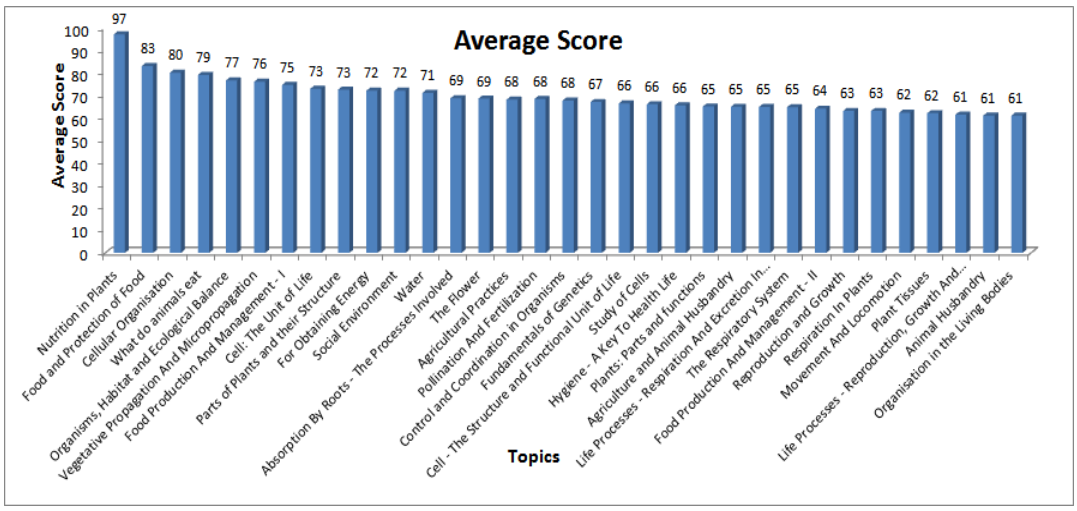

Fig. 5. Topic wise average score in biology.

3.4. Frequently Accessed Sequence of Topics

3.4.1 Inference

This analysis as shown in Fig. 6 helps in arriving at sequence of topics (Class-> Subject->Chapter-> Topic) is most preferred and least preferred by a specific learner.

4. Discussions

This analysis of web log data shows how the e-content was accessed. Using this graph, the educationist has an overview of the access made by learners in the course with a clear identification of patterns and trends. Topic wise frequent access patterns indicate what topics stimulated interest, in Biology, by learners of a particular grade. The frequently accessed topics could be due to their magnitude of complexity, or are of interest or the information provided on these topics might be more relevant to them. The analysis also showed least accessed topics, reflecting either learner's previous comfort level/expertise in the above resources or they may have found the information in these topics irrelevant to them. Using this information, educationist can detect topics which are not explained well, for example, those with a very low number of

accesses. Future research on weak content that learners don't find useful would help in arriving at steps to make e-content more useful.

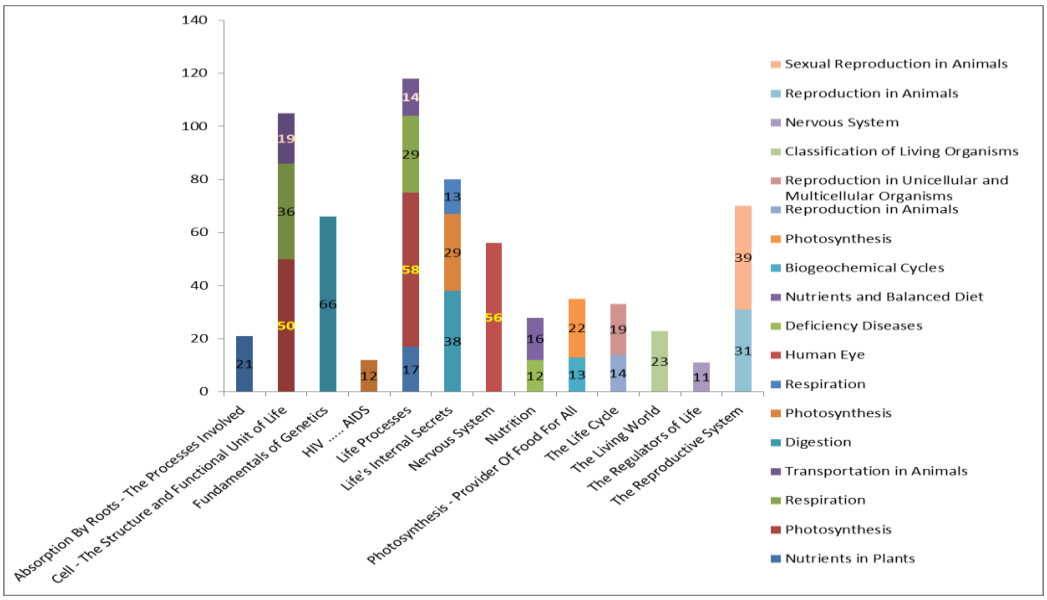

Fig. 6. Frequently accessed navigational path (class-> subject->topic-> sub-topic).

The pattern analysis also revealed that the maximum time spent in various topics by varied learners. Similarly, the topics where the least amount of time was spent could be possibly due to irrelevant content.

In the context of e-learning, we can find a meaningful hierarchy of navigation paths (Class->Subject->Chapter->Topic) taken by the learners. Mining these frequent patterns, constructed for each learner's trail, one can find more relevant hierarchy. This helps us to identify, which parts of the course/sequence of topics are used the most and least [8]. These frequent navigational patterns, among previous users, enables e-learning sites to organize the site content more efficiently or provide effective recommendations. The recommendations hence obtained are in the form of rules, where the antecedent of the rule is matched with previous learners' trail and consequent are used to recommend the links to new learner [11].

This can enable an e-learning site to reorganize the site content more efficiently or to recommend additional topics. It helps in identifying those interaction sequences which indicate improvement and to explore how to reorganize the placement of material and assessments on the basis of usage and performance data.

5. Future Scope

Although, interesting patterns could be mined, it requires future work such as using web mining methods to incorporate previous learners' feedback for making recommendations. This further entails how to identify those who could benefit from feedback and how to decide what feedback could be most effective .Further study can be done to arrive at better visualizations. This information on actual web usage by a learner can help in adapting a website to suit another similar potential learner.

References

[1] Nadia, Y., & Nisreen, A.-B. (2013). The impact of changing technology: The case of e-learning. *Contemporary Issues in Education Research*, 173-180.

[2] Radenkovi, B., Despotovi, M., Bogdanovi, Z., & Bara, D. (2006). Creating adaptive environment for e-learning courses. *JIOS, 33(1).*

[3] Miller, J. (2005). Usability in e-learning.
[4] Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Evolution of Teaching and Learning Paradigms in Intelligent Environment, 62*, 183-221.
[5] Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM, 43(8),* 142–151.
[6] Etzioni. O. (1996). The world wide web: Quagmire or gold mine. *Communications of the ACM,* 65-68.
[7] Pamnani, R., & Chawan, P. (2010). Web Usage mining: a research area in web mining. *Proceedings of ISCET 2010*. Punjab, India.
[8] Mahajan, R., Sodhi, J., & Mahajan, V. (2014). Usage pattern discovery from a web log in an Indian e-learning site. *Education and Information Technologies*.
[9] Hu, W.-C., Zong, X., Lee, C.-W., & Yeh, J.-H. (2003). World wide web usage mining systems and technologies. *Journal on Systemics, Cybernetics and Informatics, 1(4)*.
[10] Cotton, D., & Grestya, K. (2007). The rhetoric and reality of e learning: Using the think-aloud method to evaluate an online resource. *Assessment & Evaluation in Higher Education*, 583-600.
[11] Mahajan, R., Sodhi, J., & Mahajan, V. (2012). Mining user access patterns efficiently for adaptive e-learning environment. *International Journal of e-Education, e-Business, e-Management and e-Learning, 2*, 277-279.

Renuka Mahajan is an assistant professor of computer science at Amity University, Uttar Pradesh. She is currently pursuing her PhD. (CS&E) from AUUP. Her current research interests include educational data mining and e-CRM. She has published in the Springer Journal EAIT, Inderscience IJIL and IJLEG, IJEEEE and few other professional conference proceedings.

J. S. Sodhi is a CIO, Head-IT designated as an assistant vice president at AKC Data Systems Pvt. Ltd. He received his doctorate from Amity University in information security. He has participated as a distinguished speaker at various national & international conferences, published in various prestigious journals and given guest lectures to management students.

Vishal Mahajan is a group project manager at HCL technologies, Noida. He is currently pursuing PhD. program from MS Sukhadia University, Udaipur. His current research interests include data mining and telecom. He has published in the Springer Journal EAIT, IJEEEE, Inderscience IJIL and IJLEG and few other professional conference proceedings.