Educational Web Mining System Based on Result Cache Method for Information Retrieval

N. P. Joshi, R. B. Kulkarni

Abstract — There were times in the past when it seems harder to find the information on specific topics and now, the same task is just a one click away. In current times, a huge pool of information is available on different topics on World Wide Web (WWW) and the task of finding a specific one becomes a bit tricky. Different techniques are out there for information searching and retrieval, from which one can choose an efficient technique called web mining. Web mining is the technique that is used for the extraction of useful information from across the available information. There are different such a sub-techniques through which a web mining can be implemented but has some challenges and issues such as network speed, longer fetching time, content availability, lack of information relevance etc. The paper presents the methodology that tries to avoid or minimize the above mentioned problems by the means of result cache approach that reduces the fetching time, increase the availability of the information resource and provides the closely accurate resources to its users.

Index Terms—Educational Web Mining, Web mining, Information retrieval, Result cache,XML

I. INTRODUCTION

Web mining research in the field of Web Education System has become the hot spot[6].In recent years various proposes for web education system[WES] and/or web education mining[WEM] are reported .Therefore it appears that efficacy of these models are required to be estimated to build an efficient web education system [WES] , aimed at a specific requirement . It could be seen that, existing WES, which provides network education environment, not only allows learners to conveniently learn via the network, but also realizes the integration between teaching progress and learning resources. However, for the phenomenon of data explosion with poor knowledge for network education, the existing WES usually has several problems in the functions, such as weak usability of personalized learning , the low efficiency of intelligence learning, and the poor and obsolete online learning resources[1][4]. Existing WES possess lack of interactivity, cooperation and low degree of resource sharing. Thus, learners have to spend a lot of time and energy to grasp useful learning information, and it may degrade the self-confidence of an individual .Various other features of good quality WES and WEM are already discussed by different authors [1].

Manuscript Received on January 2015.

Miss. N. P. Joshi, Student, Master of Engineering, Department of Computer Science and Engineering, Solapur University, Walchand Institute of Technology, Solapur, Maharashtra, India.

Dr. Raj B. Kulkarni, Assoc. Prof., Department of Computer and Science Engineering, Walchand Institute of Technology, Solapur, Maharshtra, India.

The present paper aims at engineering or a science student as a group of learners. It could be seen that a huge amount of literature is available on web for the targeted group of learners. However the web information lacks specifications about its contents and therefore learners waste their time in surfing the internet until they find the information of their meet [2]. Another problem that exist with the web content is its unstructured format. The web content frequently available is in HTML form, which is in some context an semi- tructured format and while making it available to the users, the need stand for the data cleaning and bring the data in to proper structured format as well. To compromise with HTML we use XML language. XML is a language for defining markup languages [3]. The Advantages of XML are already discussed by Shengjian Liu et-al. So, there exist a need of a system aiming at effective enhancement of intelligent education service. Education Web Mining [EWM] provides a solution for solving the bottlenecks of existing WES [7]. The present system works on the principle of xml conversion, storage and retrieval of the html documented resources of the WWW. Here, the overall functioning of the system involves the activity of fetching web content based on certain frequent keywords and are preprocessed and brought into the standard structured format of XML. This forms the offline XML database of the system consisting of set of XML files. The user entered keyword is then searched for its existence or occurrence in the content nodes of the each XML document and the results are then given to the users based on the occurrence frequency of the entered keyword. Learners can perform search operation and interact with each other [4]. Top 10 such a results with highest frequency among them are displayed to the user. At present the WEM component is designed and implemented in an isolated environment, however the implementation is extendible to be form as a network system or WEM component. The figure above (see fig:1) presents the pictorial representation of the system. In this at the extreme top-left hand side, an admin of the system is present. As mentioned above the admin of the system is responsible for various responsibilities. The main responsibility of the system admin is to find the best relevant educational resources, regarding the educational need of the students of an institution or an organization. He/she personally queries the search engines against various keywords and finds the web resources. It is also difficult task for the admin to find such resources, but as it is manual that results more accuracy than an auto-generated system. Because, the auto generated system may or may not check accurately for content relevance.

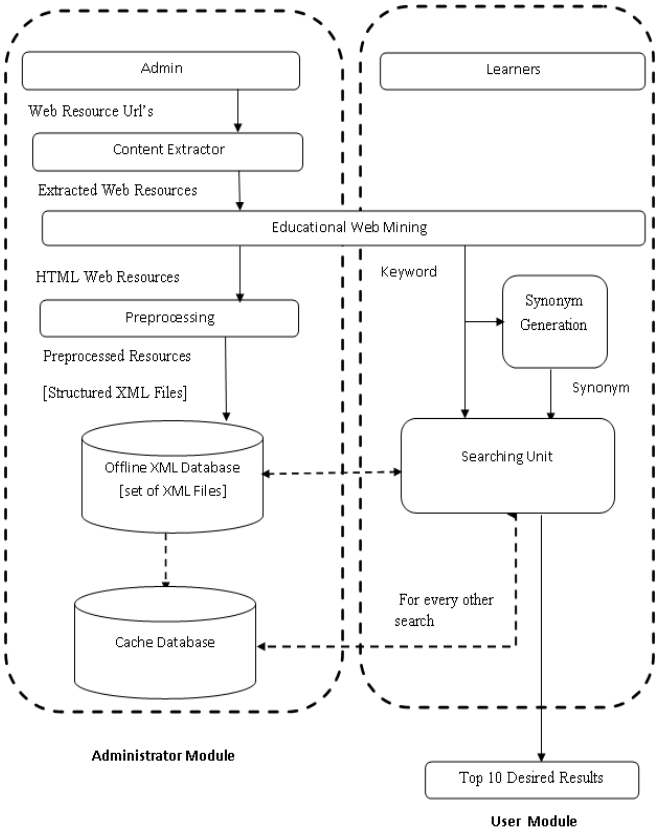

Figure 1

Another specialty of the admin module is, there are more than one administrators according to the different branches or the streams of the institution. The, administrator then collects only the URL's of such useful relevant web resources and then it gives it to the system. The system then fetches the content of those web resources with the help of content extractor. At present many websites are built with HTML, which is difficult to achieve real effective and accurate web mining[5]. As the collected data is a web resource then it might be in structured or an unstructured form. The prediction about structure can't be done dynamically. Hence, the detection can be done about the document structure and if it is semi-structured one, then it is converted in to a structured format by the process of tokenization using HTML to XML convertor in the preprocessing unit and is stored in an XML file. Set of such relevant XML files are then stored in to the database i.e. in offline XML database. Also, as there are many more resources stored into the data base, it requires an efficient fetching system for the desired results to be fetched for. So, for efficient fetching it introduces a method of result cache. For the first time if a keyword appears, the linear searching of the files takes place for its occurrences and the results are then returned and are cached at that time. Then onwards for every next searching instance of that keyword, the results are then retrieved from the result cache instead of linear searching again. This reduces the fetching time up to some lager extent. In this a searching unit works of user entered keyword .At the same time the synonym generation is also provided, so that if the word entered by the user doesn't retrieves appropriate results, then one can select the synonym as the keyword for the searching and can find the semantically same results. The paper continues with the succeeding sections of literature review, methodology, result and discussion and finally concludes in the section of conclusion.

II. Related Work

Previous work done in the any filed is always serves as a guiding path in the field of research for the novice researchers. The methods previously implemented always shows the ways in which one should or shouldn't move along and hence the literature is very important thing to be taken in to consideration and should always be appreciated as well. Hence, this section describes the literature work previously done by different authors in the field of WES or EWM (Educational Web Mining). LIU Shengjian , WU Xiaoning discusses majour characteristics of web mining technology by analyzing the current problems of IT –based education platform (ITEP).The experimental result of proposed architecture design is achievable. The target of the paper is solve the existing problems of ITEP such as weak data retrieval system , low degree resource sharing and less interactive. The further study requires adding intelligent system ITEP. The paper not only deals with technical issues of ITEP but also consolidation of advance technology through ITEP [4]. Qin Wei has analyses several issues of existing web-based teaching platforms. The existing platforms there are some problems in functioning. So paper discusses the architecture of teaching platform and bringing it together with the current popular technology called data mining. This architecture not only meet the requirement of study through providing personalized recommendations and resources, but also track and manage the status of learners. Excavating information that helps learners personalize the learning from teaching resources or vast amounts of irrelevant database records has become an important application of data mining in the field of education. The paper provides the solutions in terms of different modules. One of the modules called functional module is important one. It consists of different tasks in regard with student as well as teacher. It makes the availability of different resources to the both students and teachers. Also, it groups the students, and forms their clusters according to their interests. Based on these results, it provides the resources to them. Also, it checks for the progress of the students, by caching the queries asked by them and grouping the student with the same queries together. Then, it prepares the inference that the topic on which the queries are asked most is to be repeated to strengthen the concepts and gives such remarks to the respected teachers. Also, with the help of focused crawler module, it downloads the resources automatically and are then classified according to the students need. These resources are then registered with the libraries called as learning resource library. There is also an adaptive test module that takes the tests and provides intelligent tips to the students for their upliftment [2]. Olivier Liechti, Mark Sifer, Tadao Ichikawa proposes framework which builds around Structured Graph Format (SGF) which supports the description of Web sites structures .SGF is nothing but efficient generation of metadata..It divides the site in to sets of nodes. The nod elements are formed from the links. There are two types of links are considered called as hierarchical links and associative links. The links that follows or extends one from the other, on two different levels can be called as hierarchical links while the links extending one from the other on the same level can be called as associative links. Mostly the

workings consist of four modules called as SGF Specifications, providers, generators, and applications. Specification uses SGF Data Type Definition (DTD's) to define the standard way of information exchange from the other modules. Providers are the web sites that publish documents on the above specification basis. Generators are the are the agents that supports creation of SGF documents. The last ones are applications that that fetch the metadata published by the \web sites and process it for any desired purpose. In this paper mainly two SGF applications and three methods for generating SGF are defined. SGF framework is collection of interoperable software .The framework incorporates SGF, an XML based format, ii) applications that use SGF metadata for some purpose and iii) methods and agents that support the generation of SGF metadata. The paper concludes with the three methods of preparing metadata with comparison. The first is applications which support user navigation by dynamically generating interactive site maps. The second application is used to monitor activity occurring on a site. The experience with the SGF framework is results in good results and one drawback with this framework that it requires site designer which uses authoring environment. The alternative solution is that crawler can be used to avoid such problem [3].

III. METHODOLOGY

The methodology consist of two phases, one is related with data fetching based on Uniform Resource Located (URL) from various educational websites on different topics and the storage of it. The other phase does the actual work of finding resources offline and making it available for the students based on their selected keywords. Following fig. shows the diagrammatic representation and the detailed description about the work flow and the terminologies:

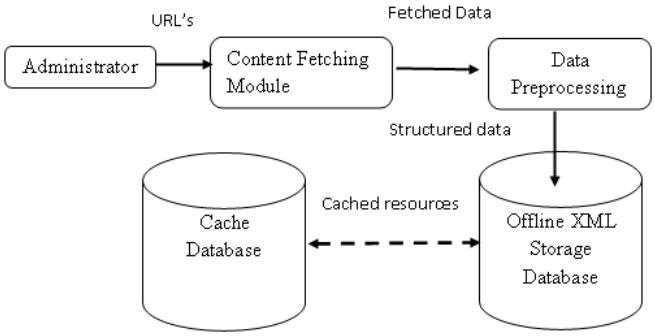

Figure 2: First phase

A. *Administrator:*

This module allows an administrator(a person) or any responsible person that concerns with the data collection regarding the syllabus of the student. The person analyses the syllabus of the students of different stream and years and tries to collect the resources from the web and stores their bookmarks in forms of URL's. It supplies those URL's to the Content Fetching Module, which in turn, fetch the content of the respected URL. Fetching data from URL's is all automated task done by the system. Another task that concerns with the administrator module is to cache the results of all the uncached keywords, so as to prepare the cached database. The content so fetched guarantee's the reliability, relevancy and accuracy as well. Also, administrator can see the list of all those uncached keywords waiting for caching, hence no need to remember those words. Only work to do is, to select the uncached word from the list and perform caching.

B. *Content Fetching Module:*

This module takes the respected URL's from the administrator module and fetch the content of that URL's. The module uses HTML parser, to parse the data from the web page of the given URL. The fetched content is then given to the next module called Offline Storage, where each data retrieved so far is stored in an xml file.

C. *Data Preprocessing:*

As discussed earlier the web content is always been a problem regarding its structure and hence to be required to undergo the process such as data cleaning, and data preprocessing to bring it to the standard structured format. So, Data preprocessing involves a check of structured or semi-structured data, and conversion of semi-structured data in to a structured format. XML is used to convert semi-structured data to well structured data [8]. This task is carried out by using HTML to XML convertor. The input is all the HTML web content or resources to this tool and the output is structured XML resources. Also, for standardization the generated XML files are checked for XML validation and if not then are converted into standard validated XML files. There are some files that only contain a line or so that is fetched due to some errors are removed here and are re-fetched and added to databases after preprocessing to form Offline XML Storage Databases.

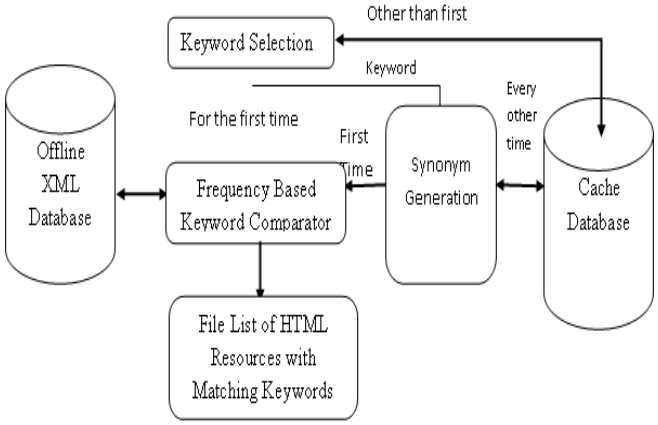

Figure 3: Second phase

D. *Offline Storage XML Database:*

This module acts as a backbone for all the functions of data storage or retrieval. All the XML files generated in the previous phase are stored here as a database. This is the biggest phase where, once the data is stored in an offline form, there is no need to go online each time to fetch the desired static content and again doing which may or may not be useful in terms of relevancy and reliability as well. The main motive of storage in XML, describes from its properties which are simplicity, ready to use, extensible and very important it separates content from presentation. So, content separation from presentation helps more easily in storage purposes rather than including the design stuff. Also, it is easy format for the data retrieval. In the retrieval process, the data from the nodes of the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

XML tags are extracted and is then used for the comparison of the keyword occurrences. The updation of this database is also done periodically for searching new resources and adding it to the databases.

E. Frequency Based Keyword Comparator:

As shown in the fig 2, this module takes two inputs, one is user selected keyword and other is Offline XML Storage Database. Here, the user is the student looking for its desired data stuff and hence enter the keyword. The keyword entered is then given internally to the Frequency Generator and Comparator algorithm, which in turn, then extracts the html pages based on the frequency of the keyword within existing XML data storage from the previous phase. The frequency calculation and the algorithm used is as follows:

Frequency of keyword for a single file =
Total no. of occurrences of the word within the file

Frequency Calculation Algorithm:

1. Select Keyword
2. Find Total no. of Occurrences of the keyword within each file and store it in hash map
3. Compare the Occurrences with each other to find the top 10 figures out of it.
4. Extract their filename
5. Retrieve from the offline database the files matching those file names
6. Convert those files to HTML format and present it to the user

So, each and every file from the offline database is then scanned for the occurrence of the selected keyword and the count of each file is maintained. The counts are then compared with each other to get the top 10 counts of occurrences and are then presented to the user in the form of list as HTML pages.

6.List of HTML Resources with Matching Keywords:

As mentioned above, the files are then retrieved based on the keyword comparison, and the list of HTML pages of desired keyword selection is presented to the user, by clicking on which the user is redirected to the clicked resource.

IV. RESULT ANALYSIS

Result Analysis is such an important phase of the research work that directly projects some conclusions from the methodologies one has worked for. It brings many more conclusive points whether in full or in partial depending upon the status of the work, in to the picture. So, the following are some results and the discussion about the system performance based on the above mentioned methodologies.

A. Dataset:

For every result to be corrective, the term called as "dataset" plays an important role. Also many of the researchers believe that the success of the entire research in the completed format totally depends upon the proper dataset. That means if no proper dataset used, then one may or may not get accurate results up to the satisfaction, in reverse to that, if a proper dataset is used then it is more than enough to measure the accurate success or failure of the research. So, the dataset considered here in this paper consist of in total 300 web page textual resources, varied in size according to content. All the 300 URL's are fetched for educational data and they are converted to the XML files. Each and every file irrespective of its size is then scanned for having presence of desired keyword in it. In total, 15000 lines are scanned for checking its appearance in each file. All these files are taken from the valid sources on the web, and are checked for their relevancy about the education. In total 4 students are taken into consideration for this study, and different keywords entered by them are stored in to the database and are then used to find the relevant document according to the need. In total the system is tested against 20 keywords for their search results each of 5 entered by a single student.

B. Experimentation results:

The result of the experimentation is tested firmly on the basis of the efficiency in terms of performance.

The result analysis here based on three measuring parameters as follows:

1) Fetching time of result extraction

As already discussed, fetching a limited amount of data from a pool has always been a challenge for the branch of mining. The time parameter here plays a very important role. The expectation is maximum no. of relevant result are fetched in minimum amount of time. So, on this measure of performance evaluation, the system performs well in both the case of linear searching method i.e. the method that checks for the first time the occurrence of the keyword and the other one of caching. But, is done the comparison the caching fetch time is more efficient that the linear one. The graphs and the readings are shown below.

Table 1: Keywords vs. Linear Time

No	Time	Linear Time
1	Binary	2
2	Education	2
3	Fuzzy	3
4	Data	6
5	Database	4
6	Information	1
7	Cobol	3

Also, when it comes to linear searching method it proves to be efficient, as the file size becoming the important factor there for the retrieval.

2) Method of searching:

This being a parameter, it is found that, amongst the two methods of linear search and the cache method, the latter is proved more efficient as it is pre-fetched entries for a single time while in the former each time the entire fetching is repeated on the re-occurrence of the keyword. On the basis of time both the methods are compared with their respected readings below:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

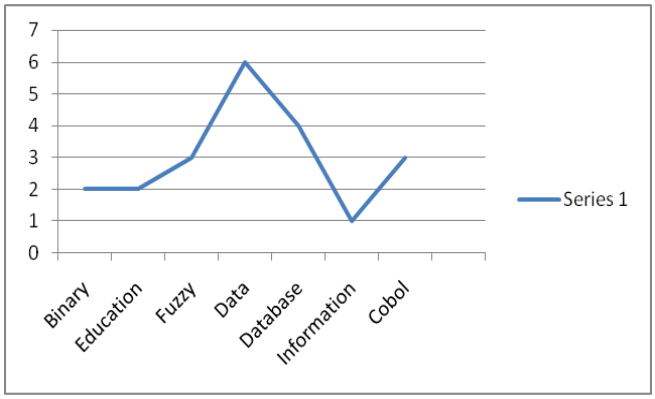

Figure 4: Graph 1: Keyword verses Linear Time

1. Keyword verses Cache Time fetch :

Table 2: Keyword verses Cache Time

No	Time	Cache Time
1	Binary	1
2	Education	1
3	Fuzzy	3
4	Data	1
5	Database	1
6	Information	1
7	Cobol	1

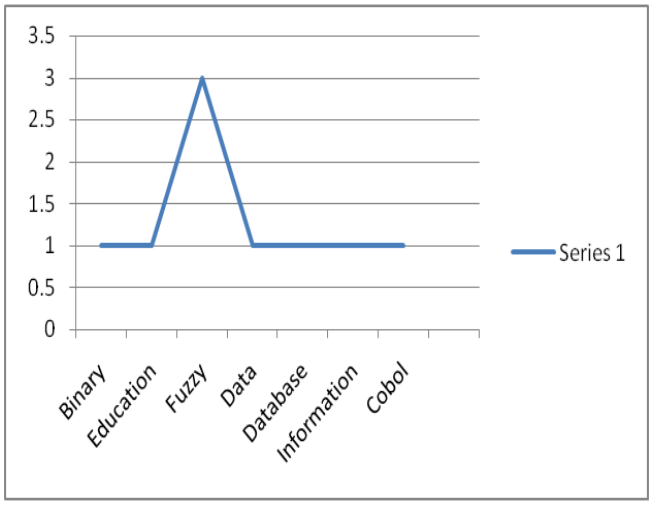

Figure 5: Graph keyword verses Cache Time

So, from the above graph it is seen that, as the file selection for retrieval is dependent of user selected keyword. Also, file with the greater file size may contain the least relevant data and the reverse is also possible. In figure above (see figure 4), it shows the graph of keywords vs. linear time, showing efficiency more greatly than the figure 5 i.e. with cache time. Also, the figure below shows the pie-chart representation of the keyword vs. % fetched results.

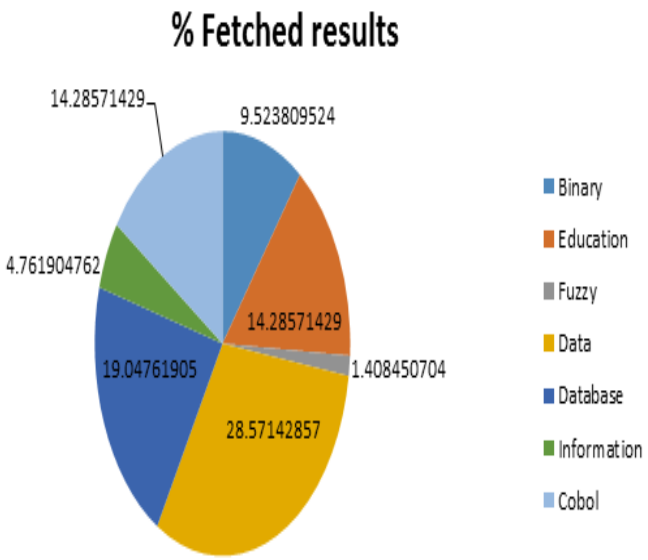

Figure 5: Pie chart-keywords vs. % of fetched results

Note: Pi-chart drawn above changes according to change in data.

V. CONCLUSION

There is always a need for the efficient ideas to come out for the upliftment of the every field, and so does for the field of the EWM. Earlier concepts of EWM involve searching the resources online, mine them and put them in front of users. The resources provided by the EWM sometimes may be useful sometimes may not. The problem arises here only, of providing a good relevant and accurate educational web resource to the users. The system mentioned above tries to minimize or remove the issues of EWM such as resource availability, relevancy, accuracy etc. The system or its administrator mines for the desired data on to the WWW and fetch the data by providing its URL's to the system. The fetched data is then tokenized in the XML node formats and stored on to the system. The user can search the stored data offline, by providing his/her keyword to the system. The system then uses counting algorithm, counting no. of occurrences of a keyword in to each file. Based on this no. the comparison of counts is done and top 10 such a results are displayed to the users that matches with the keyword provided. The system uses hash map concept, by storing the keyword and their counts in to the key and a value pair that makes system very efficient in time complexities. Also, the concept of result cache makes it easy the task of result fetching from the available pool of data reducing the fetching time. Also, the system assures on an average 6-7 results to its users which are most relevant .As the system is offline, no need for the internet connection and hence avoids the overhead of message exchanges during communication. Also, as it is applied for the purpose of education, i.e. institution or an organization, it restricts the no. of users increasing availability of offline web resources and each resource is checked against its relevancy, that issue is also resolved.

REFERENCES

[1] Liu, Shengjian, and Peiyuan Liu. "Research of Educational Web mining based on XML." Computer

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Science & Education (ICCSE), 2012 7th International Conference on. IEEE, 2012.
[2] Wei, Qin. "Research and Design of Web-based Teaching Platform." 2010 Second International Workshop on Education Technology and Computer Science. Vol. 3. 2010.
[3] Liechti, Olivier, Mark Sifer, and Tadao Ichikawa. "A metadata based framework for extracting and using Web sites structures." Multimedia Computing and Systems, 1999. IEEE International Conference on. Vol. 2. IEEE, 1999.
[4] Shengjian, Liu, and Wu Xiaoning. "Architecture design of IT education platform based on web mining." Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on. Vol. 4. IEEE, 2011.
[5] Lan, Li, and Rong Qiao-mei. "Research of Web mining Technology based on XML." Networks Security, Wireless Communications and Trusted Computing, 2009. NSWCTC'09. International Conference on. Vol. 2. IEEE, 2009.
[6] Liu, Shengjian. "Educational Web Mining Applications in Intelligent Web-Education Systems." Information Technology, Computer Engineering and Management Sciences (ICM), 2011 International Conference on. Vol. 4. IEEE, 2011.
[7] Romero, Cristóbal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146.
[8] Kumar, M. Kiran, Shaik Rasool, and S. Jakir Ajam. "Web data mining Using XML and Agent Framework." IJCSNS 10.5 (2010): 175

N. P. Joshi, is pursuing her M.E. in Computer Science and Engineering from Walchand Institute of Technology, Solapur University, Maharashtra, India. She received her B.E. in Computer Science and Engineering from Shri Vithal Education & Research Institute, College of Engineering, Pandharpur,Dist. Solapur, Maharashtra, India. Her areas of interest includes data mining, web mining and usage of web mining for the purpose of educational field.

Dr. Raj B. Kulkarni, received the Ph.d from Solapur University. He is Associate Professor in Walchand Institute of Technology. His area of interest includes Restructuring and web mining. He has 8 papers published in International journals, 20 papers published in National and International Conference Proceedings. He is a CSI member and worked as Akash Workshop coordinator held by IIT, Bombay. He had attended many workshops of National and International level. He has experience of 22 years in teaching. Currently is the chairman of Board of Studies, of Solapur University, Solapur, Maharashtra, India.