ISSN:0975-9646

Offline Resource Extractor: An Efficient Resource Extraction System

N.P.Joshi
Information Technology ,WIT solapur,India

***Abstract*— Data and information are the two most discussed terms in every filed. The both have their own peculiarities in one or the other respect. Data is growing exponentially, while the techniques for searching the information are increasing day by day. Though there are a large set of techniques available for information search, it has been found that there is a need of a technique that will be more efficient in providing availability, accuracy and relevancy for finding information or particular resource. The proposed system aims at providing the system that facilities its users to extract the relevant searched resource on offline basis. The 'resource collector '(RC) of the system collects the resource from the World Wide Web (WWW), in prior and then store it offline on to the system database. Then, as per the user request/query, the particular resource is handed over to the user.**

Keywords*— *Educational Web Mining,Frequency count,Resource Searching,Resource Finding, Information Retrieval

I. INTRODUCTION

Searching is becoming a habitual part of the life, for any queries or questions to be solved for. Almost everybody is seen, goggling for anything for resolving their queries at least once a day, irrespective of their professions. So, while searching the things on the search engines doesn't guarantee all the times the intended information one is looking for. See the following diagram.(see Figure 1). Here, it's a structure is like cycle, where it starts from its first phase of information search. The user pass into his/hers query to the search engine and does the information searching. Today's search engines are smart enough to look after for the intended search results. But, there are lot more situations the user is facing and complaining about, the data they are looking for is either not available or is not the desired one. So, after passing through these different phases of the below cycle, they reach up to the last phase of it and last their search into yet another new search and the cycle continues. Also, another thing is it in need of an internet connection, every time a query appears, which will be a time consuming one. So, the need arises to break this cycle and brings a system that could solve this problem. The proposed system tries to solve this problem completely. The first part of the system covers resource collection and second part consist of resource delivery. The main audience for this proposed system is all the people related with the field of academic education. In the first part, the educational resources are collected by taking, syllabi of the different students of various streams in to consideration, these resources are then given to the filter and are preprocessed and are stored in to the databases. The second part consist of the delivering the intended resource to the desired user. In this, distinct algorithms are used to in order to identify the useful resources and are then given back to the users.

Fig. 1. Information Search Life Cycle

Top 10 such results are retrieved for the user entered keyword, if available. The proposed system can be studied in different various different perspectives such as follows:

A. Learner's Perspective

This perspective describes the learner approach towards the system. The learner's category can cover the student's aspect, who many times search on the internet for the educational resources. They might be looking for a part of a particular concept or a complete e-book to resolve their queries. They are also finding this difficult to search on the internet for the required resource and many times getting disappointed. So, the proposed system is very much helpful to them. They can find the educational resource they are looking for quite easily and as the resource is from their syllabi, it assures its availability as well. Also, it is more efficient and beneficial as the resource is made available offline and are authentic ones. If the resource is not available, they can report it to the RC, and can demand it for its availability as early as possible.

Essentially, the main aim is to study and analyze the entire data on the web, and extract the essential and necessary part according to the student entered queries. This facility may benefit student in searching the intended educational material such as notes, presentations, reports, technical programs for their references. Also, some reference books that are not possible to be purchase, and are available as e-books can be brought up to students.

B. Teacher's Perspective

It's becoming a need for the teachers to be in a flow of modernization, and to keep them updated according to the change as many changes occurs in to the syllabi periodically. So, like students teachers have to look for the required material of resource on to the internet. Also, they

have many important responsibilities that teachers handles and resource collection and scrutiny for the best relevant is tricky one. So, the system facilitates the teachers also to search for their required resources and get it whenever needed. They can enter the keyword and fetch their available required resources from the system without internet connection. Also, they can guide the RCs, in collecting the proper resource and resource updation as well if needed.

II. Related Work

Qin Wei has analyses several issues of existing web-based teaching platforms. So paper discusses the architecture of teaching platform and bringing it together with the current popular technology called data mining [1]. The architecture of platform combined with the current popular data mining technology. This architecture reach at the need of personalized study through providing personalized recommendations and resources as well as it track and manage the status of learners. Thus the personalized learning was brought out. 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 able to be used 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 examines 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.

Olivier Liechti, Mark Sifer, Tadao[2] Ichikawa proposes framework which builds around Structured Graph Format (SGF) which supports the description of Web sites structures. The Structured Graph Format is based on structured graphs.SGF is nothing but systematic generation of metadata. It divides the site in to sets of nodes. The node elements are formed from the links. There are two types of links considered, called as hierarchical links and associative links. Hierarchical links are simply links that follows or extends one from other and those are on different levels while the links extending one from other on same level known as associative links.

The working consists of four modules known as specification, 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 working as like web sites that publish documents on the basis of above specification. Generators are the agents that support 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 nothing but collection of interoperable software .The framework includes SGF, i) it is based on XML 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.

LIU Shengjian , WU Xiaoning[3] discusses majour characteristics of web mining technology by analyzing the current problems of IT –based education platform (ITEP).This architecture reach at the need of IT personalized education through providing personalized learning function and resources, as well as gives some solutions and way of improvement of ITEP. The experimental result designates that this architecture design is workable. The success of the educational web mining work requires future work in direction to design a intelligent educational platform.

Robert Pinter et. al.[4] discussed the recommender system in E- student web based adaptive educational application. Initially they discussed how people believe in recommendations by giving some examples of day to day life.Then they discussed how recommendations can be utilized saying it is the task of taking the information on some topic by maximum people, and delivering the offers or suggestion to the needy people having interests in that same field. Further they say, the recommendations can be mainly formed on the two bases. One is marking basis and the other one is user activity tracking. The former is the well known method that is utilized in many of the applications. In this, an application suggests the user to rate for the particular thing on certain scale says 1-10. So, then it enumerates the average mark from the obtained readings and based on that rating generates the recommendation. In the latter method, it is the system oriented process. Here, the user preferences are stored automatically by the system about some particular thing and based on those preferences the system generates the recommendations. The best example of it is page rank algorithm. i.e. the page is ranked according to how many people have visited it. In their recommendation system, they have taken 5 types of recommendations such as Number of visits, Evaluation, External Sources, Comments, Navigation Assistance.

Federico Michele Facca, and Pier Luca Lanzi[5] presented a wide survey regarding mining the interesting patterns of knowledge from web logs. They have discussed types of web mining especially as web content mining, web structure mining and web usage mining.

The data sources they have used are web servers, proxy servers, web clients. The main problem that they have stated is how to group all users' page requests i.e. paths followed by them during page navigation. This problem is tackled by using the option of cookies. Also, for data preprocessing, they have used four tasks to complete. The first is data cleaning, followed by identification and

construction of user's session, then it get backs the information of page content and structure and then have a phase of data formatting in the required format. They have also discussed several techniques such as association rules, sequential patterns and clustering.

Martin Labaj and MáriaBieliková [6] in this paper have discussed modeling parallel web browsing behavior for web based educational systems. The term parallel browsing here means, the browsing that is done by sidewise while doing some another browsing as well. They have captured the resources accessed by the students side by side while learning the particular concept. Many students have the practice of browsing the other sites for same information. The best example of it is searching a topic on google.com. There are many results that search for the single entered query. Each result is then opened in the other window and can be treated as the parallel browsing. They have done the study of parallel browsing with different approaches such as tracking the user behavior at server as well as client side. At server side the server logs are taken into consideration for one continuous session of users. At client sides, user histories such as user clicks and user caches are taken into consideration. They have considered discrete events such as page load, page unload etc. Also, they have tracked which user have opened which page and when along with different tabs.

Neelamadhab Padhy [7], et. Al, presented the survey of different data mining applications and their future scope. So, this can be then applied in case of web mining as well for resource collection purpose.

Chaware Anita [8], also described emerging trends and technologies in the field of education data mining.

III. PROPOSED WORK

Proposed work describes the actual work to be implemented and working and responsibilities of different functionalities of the system. See the following figure of Educational Resource Finder System (see Fig.2).

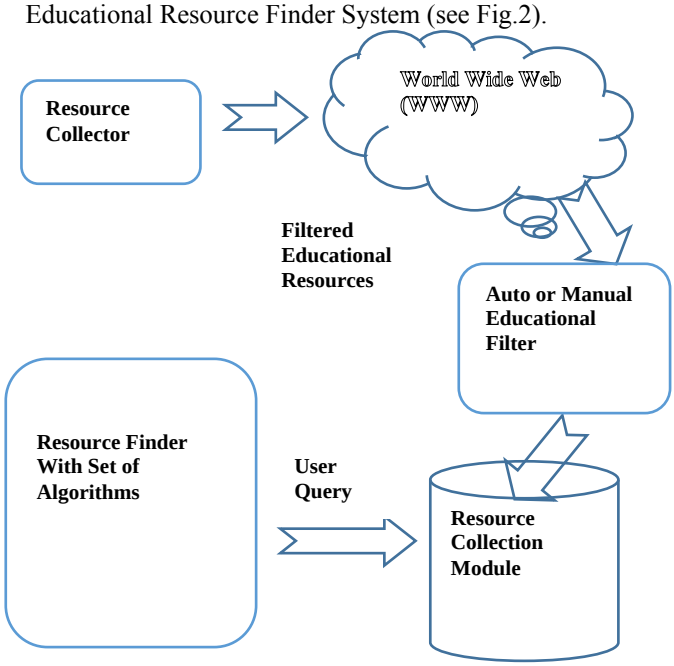

Fig. 2. Educational Resource Finder System

A. *Resource Collector(RC) :*

Resource Collection is the crucial part of this system, as ultimately they are ones that to be delivered to the end users. The focus here is to collect the proper resources that could really benefit the users and are not compromised on the account of availability, authenticity, correctness and accuracy. The one approach is to automate the process and write the code that on certain parameters can collect the resource. But, that might not work accurately and doesn't assure the relevancy as well. So, the other approach is to have a physical intervention that can manage the job of resource collection. This approach is the secure one and most promising too. For each subject a specialist teacher can collect the resources and add them to the databases. Also, as the teacher is teaching the subject, he/she can authenticate the resource as well. If it is not possible to have a single teacher for each of the subject, a domain expert can be assigned, for eg. A person having all the knowledge of networking, as a domain expert, instead of a single teacher for all n/w related subjects.

B. *Auto or Manual Educational Filter and Resource Collection:*

While collecting the resource, as authenticity is the one problem, the other one is irrelevancy. While searching on to the WWW, there might be the possibility that the too many resources that are irrelevant to the subject comes at the top and when opened for, they all disappoint the users. Also, some fake kind of results often comes in place, where the user has to navigate across different web pages and at the end doesn't find the intended resources. Another case is of ambiguity. There might be two possible meanings of the same word. For instance, a word 'JAVA' can be searched for java beans or for a programming language java. So, taking these many possibilities, into consideration, the filtering is needed to collect the proper resources as per the need. Also, the filter does another function of preprocessing. Here the textual resources collected are not in a structured format. They almost all are in HTML format. So, there exists a need for conversion from unstructured HTML format to structured XML format, to store it in to the database. So, this conversion is also carried out by using HTML to XML conversion.

Here , The specific advantages of XML those are for utilized for the proposed offline resource finder are summarized as follows.1) New tags can be created as they are required.2)Tags, attributes and element structure provide context information , opening up new possibilities for highly efficient search engines, intelligent data mining, agents, etc[11].3)XML tags describe meaning not presentation[10]. 4)XML documents can contain any possible data type — from multimedia data (image, sound, video) to active components (Java applets, ActiveX)[10].5)Reflects structure and semantics of documents → better searching and navigation.6)Improved searching of web based documents ,with huge implications for Intranets.XML can make easy learning searches .We can define XML –based metadata vocabularies for standardized tagging of learning resources. What he is supposed to do is to pose a query against each available XML document in order to extract the knowledge[9].

The XML files thus collected could be search again for validation of the keywords or any other relevant tag. This part of the resource collection is most important activity .The process of validation is an intelligent and flexible activity .The related moule could be redesigned or replace to add context specific intelligence to the offline resources , such that the offline resource finder serves for improved satisfaction.

C. Resource Finder:

This one is the important phase where the actual offline resource finding is performed, to deliver to the end user with the proper intended resources. The resource finder takes the user entered keyword as the input. Then it searches the occurrence of the keyword in to the files of the databases and then maintains these counts along with the file names. At the end of this procedure, all the counts of the keywords along with their file names are compared, and the top 10 results are then retrieved and given to the end users. Also, in case if the resource not found, then it can be searched with its synonym as well.

Count-Map Algorithm:

1. START
2. Take the input keyword
3. Search the keyword occurrence from the files in the databases
4. Maintain the keyword count i.e. keyword frequency.
5. Keyword Frequency= ∑keyword occurrence per file.
6. Map keyword frequency to the proper filename
7. Compare the counts at the end and return top 10 results

IV. CONCLUSION

The proposed system is the one that facilitates the proper resource finding and delivering it to its intended end users. As the system is simple and works on simple count-map algorithm, very much fast to operate with. Also, as the resources are authenticated and accurate enough, to ensure the availability as well. No internet connection is required, as the system works offline and hence will be available anytime. No problems such as network overhead, failure etc. occurs. The future enhancement can be done by doing improvement in resource retrieving methods i.e. by using techniques such as page ranking, caching etc.

REFERENCES

[1] Qin Wei,"Research and Design of Web-based Teaching Platform",School of Information and Electronic Engineering.

[2] Olivier Liechti,Mark Sifer, Tadao ,"A Metadata Based Framework for Extracting and Using Web Sites Structures",IchikawaInformation Systems Laboratory1-4-1 KagamiyamaHigashi-Hiroshima 739, Japan

[3] LIU ShengjianWU Xiaoning, "Architecture Design of IT Education Platform Based on Web Mining"©2011 IEEE.

[4] Pinter, Robert, et al. "Recommender System in E-student web-based adaptive educational hypermedia system." MIPRO, 2012 Proceedings of the 35th International Convention. IEEE, 2012.

[5] Facca, Federico Michele, and Pier Luca Lanzi. "Mining interesting knowledge from weblogs: a survey." Data & Knowledge Engineering 53.3 (2005): 225-241.

[6] Labaj, Martin, and Mária Bieliková. "Modeling parallel web browsing behavior for web-based educational systems." 2012 IEEE 10th International Conference on Emerging eLearning Technologies and Applications (ICETA). 2012.

[7] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope "

[8] Chaware, Anita. "Educational Data Mining: An Emerging Trends in Education." International Journal of Advanced Research in Computer Science 2.6 (2011).

[9] shengjian Liu ,Peiyuan Liu "Reserch of Educational Web Mining based on XML " ,The 7th inernational conference on Computer Science & Education (ICCSE 2012) July 14-17,2012.Melbourne ,Australia.

[10] B. Tommie Usdin "How and Why Are Companies Using XML?" Mulberry Technologies, Inc. http://www.mulberrytech.com/

[11] By Lucjan Pawlowski, Marzenna R. Dudzinska, Artur Pawlowski "Environmental Engineering Studies: Polish Research on the Way to the EU"

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.