A Web Mining Approach for Personalized E-Learning System

Manasi Chakurkar[1]
PG student, Department Of Computer Engineering
MIT, Pune, INDIA

Prof.Deepa Adiga[2]
Assistant Professor, Department Of Computer Engineering,
MIT, Pune INDIA

Abstract—**The Web Mining plays a very important role for the E-learning systems. In personalized E-Learning system, user customize the learning environment based on personal choices. In a general search process ,a hyperlink which is having maximum number of hits will get displayed first . For making a personalized system history of every user need to be saved in the form of user logs. In this paper we present a architecture with the use of Web mining for Web personalization. The proposed system provides a new approach with combination of web usage mining, HITS algorithm and web content mining. It combines hits results on user logs and web page contents with a clustering algorithm called as Lingo clustering algorithm. This proposed system with combined approach gives a better performance than a usage based system. Further the results are computed according to matrices computed from previous and proposed method.**

Keywords—Web usage mining; web content mining; web personalization; e-learning system, Lingo; HITS

I. Introduction

E-learning is a form of electronic teaching that enables people to learn anytime and anywhere. The objective of an online personalization system is to provide users with the information they require, without asking them explicitly [1]. Web personalization is outlined as any action that adapts the data or services provided by a website to the requirements of a selected user or a collection of users, taking advantage of the information gained from the users guidance behavior and individual interests.

The prominent characteristic of E-learning system is that the students become self-learners and explorers of knowledge. In learning process, learner is the main body but not passive recipient. E-learning system is built on web service related technology and provides all kinds of learning ways to students to realize real-time, interactive and cooperative learning at different places. It should be able to find the individual differences of learners and construct personalized learning environment to meet their individual needs. Therefore, more and more researchers have done much on personalizing learning system [1].

A core part of the personalization method is generation of interested information to users. It's strictly supported as certain patterns, and ensuing probabilities.

Usually used user models are rather simplistic, representing the user as a vector of ratings or employing a set of keywords. Even wherever additional multi-dimensional information has been accessible, like once assembling implicit measures of interest, the information has historically been mapped onto a single dimension, within the sort of ratings. In particular profiles commonly used lack in their ability to model user context and dynamics. Users rate completely different things for various reasons and under completely different contexts. The user interests and desires amendment with time. Distinguishing these changes and adapting to them may be a key goal of personalization. We recommend that the personalization process be taken to a new level, where the user doesn't to be actively attached the personalization method. All that the user has to do is to register to the system, create own profile and once the user logs onto E-learning system, the browser checks for that profile file as it checks for the cookies. The profile file describes the user's area of interest. Since the profile file is in standardized format, the contents are provided according to the profile file. This is able to enhance the user's personalization method while not their active involvement. The search component of personalized E-learning systems is as shown in figure 1 [8].

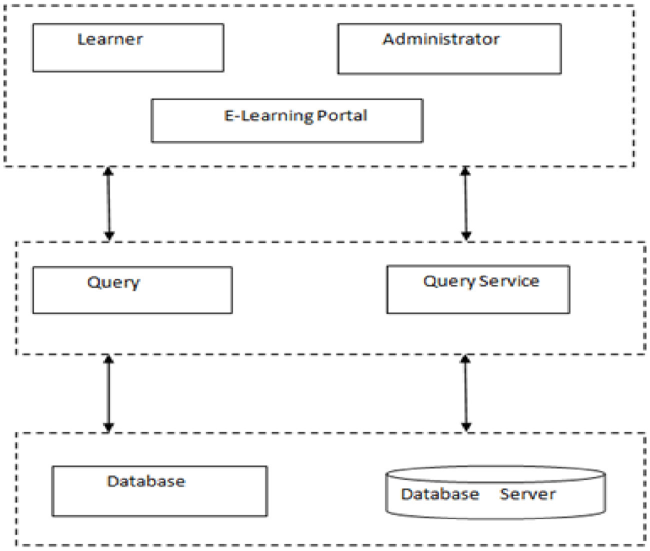


Fig. 1. Search Component Of Service Oriented Reference Architecture For Personalized E-Learning System

The Fig.1 represents search component of service oriented reference architecture for personalized E-Learning System. In e-learning systems, the challenges are increase of complexity and more interoperability between systems in distributed environment. The lacking is reference architecture in which by reusing web services, reusing learning objects etc. A service-oriented reference architecture describes essence of software architecture and the most significant and relevant aspects. [8]

E-learning portal applications need to present data in a tabular format to the users. The functionality provided by the search component is dynamic query generation based on user input, sort order, joins, etc. There are multiple ways to invoke the query service. All are supported by this protocol. Based on the service description and the protocol, new client interfaces can be easily built according to users' preferences. Learners want search the learning object related to Computer Science, then they have to type 'Computer Science' .By activating the link on the query interface, it obtains the extra services exposed by other providers. [8]

There are variations between layout customization and personalization. In customization, the location is often adjusted to each user's preferences concerning its structure and presentation. Whenever a registered user logs in, his/her customized home page is loaded. This method is performed either manually or semi-automatically. In personalization system advises in [1], modifications regarding the content or maybe the structure of an online web site are performed dynamically.

Therefore a well-designed customized distance education system might have the subsequent characteristics: i)It will find out the learners' study interest, access habits, learning orientation ii) It will change the location map and customize the learners' interface dynamically consistent with the access log iii)It will suggest learning resources by analyzing the leaner’s interest and learning process iv)It will recommend interested information v) will provide suggestions to assist him/her to regulate teaching set up, teaching model and teaching ways[1]. Principal parts of Web personalization system are firstly, the categorization and preprocessing of Web knowledge, secondly, the extraction of correlations between and across different forms of such knowledge, and finally, the determination of the actions that ought to be counseled by such a personalization system.

This survey of papers is organized as follows. In next section II we are presenting the literature survey over the web mining approach over personalize E-learning system. In section III, the proposed approach and its system block diagram is depicted. In section IV we are presenting the current state of implementation and results achieved. Finally conclusion and future work is predicted in section V.

II. Literature Survey

In this section we are presenting the various approaches those are presented to resolve the web mining approach over personalization of E-learning system.

A. *Web mining techniques*

Web Usage Mining: Web Usage Mining is the type of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web based applications. Usage data captures the identity of Web users along with their browsing behavior at a Web site.

Some of the typical usage data collected at a Web site called user logs include IP addresses, topic Id, blog Id and access time of the users.

Web Content Mining**:** Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts of a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables [9].

A survey of the use of the Web-mining for Web personalization is necessary. More specifically, they introduce the modules that comprise a Web personalization system, emphasizing on the Web usage mining module. A review of the most common methods that are used as well as technical issues which occurred is given, along with the brief overview of the most popular tools and applications available from software vendors. Moreover, the most important research initiatives in Web usage mining and personalization area are presented. But this is not as much effective method for personalization [4].

B. *HITS Algorithm for Detecting Web Communities*

HITS (Hyperlink-Induced Topic Search) algorithm, which capitalizes on hyperlinks to extract topic-bound communities of web pages. Despite its theoretically sound foundations, they observed HITS algorithm failed in real applications. In order to understand this problem, Saeko and others developed a visualization tool LinkViewer, which graphically presents the extraction process. [3]This tool helps to reveal that a large and densely linked set of unrelated Web pages in the base set impeded the extraction. These pages were obtained when the root set was expanding into the base set. As remedies for this topic drift problem, prior studies applied textual analysis method. On the other hand, They propose two methods which utilize only the structural information of the Web: 1) The projection method, which projects eigenvectors on the root subspace, so that most elements in the root set will be relevant to the original topic, and 2) The base-set downsizing method, which filters out the pages without links to multiple pages in the root set. These methods are shown to be as robust for broader types of topics and low in computation cost.

C. *Clustering Algorithms - K-means, Suffix Tree and LINGO*

The performance of the web search engines could be improved by properly clustering the search result documents. Most of the users are not capable to give the appropriate query to get what exactly they wanted to retrieve. So the search engine will retrieve a massive list of data, which are ranked by the page rank algorithm or relevancy algorithm or human judgment algorithm. The user will always find to self with the unrelated information related to the search due to the ambiguity in the query by the user. Evaluating the performance of the clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm. In this paper comparative analysis is done on three search results of clustering algorithms to study the performance enhancement in the web search engine[5].

D. *Concept-Driven Algorithm for Clustering Search Results*

Search-results clustering aims to present information about the matching documents. It's like taking a step backward to grasp a bigger picture. They consider no longer care about individual documents, but about some underlying semantic structure capable of explaining why these documents constitute a good result to the query. To find this structure, the Lingo algorithm is used. The Lingo algorithm combines common phrases discovery with latent semantic indexing techniques to separate search results into meaningful groups. It looks for meaningful phrases used as cluster labels and assigns documents to the labels to form groups [6].

III. Proposed Approach Framework and Design

A. *Problem Definition*

When the intention behind the search query is not clear, user will get large number of results in return. The user need to be swift through long list of results to find the result that suits his information need. Hence Personalized E-learning system is required to provide interested information to user by web mining technique [7].

B. *Existing System based on Usage mining*

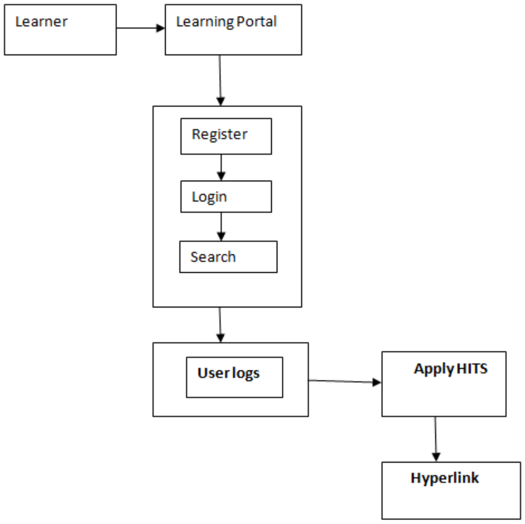


Fig. 2. Existing system based on user logs

Fig. 2 represents existing system which is based on web usage mining. In existing system once user log on to the system, and make search for a particular topic, user logs are saved. On the user logs , HITS algorithm is applied to provide the output in the form of hyperlink. Limitation of this type of system is that web page contents are not considered while giving output to the user. As a result most popular hyperlink with maximum hits will be considered, but that link may have contents which are not related to user's interest.

C. *Proposed Architecture*

The objective of this approach is to provide users with the information they want or need, without expecting from them to ask for explicitly [1]. To overcome the above stated problem, we propose an architecture, which combines web usage mining and web content mining techniques. Fig. 3 represents proposed architecture for personalized E-learning system combining HITS algorithm on user logs and Lingo clustering algorithm.

Admin is responsible for forming a standard dataset including learning objects. This dataset is preprocessed by removing stop words and stemming. A new user first makes registration to the learning portal. When user login to the system using his own user name and password and search for a particular topic, at server side user logs are stored for that search. Then HITS algorithm is applied to those logs to increase the weight of that log. For proposed approach, preprocessed data is given to content mining using Lingo clustering algorithm. Clusters are formed from preprocessed data. Final results are calculated by combing user logs, Hits algorithm and clustering results.

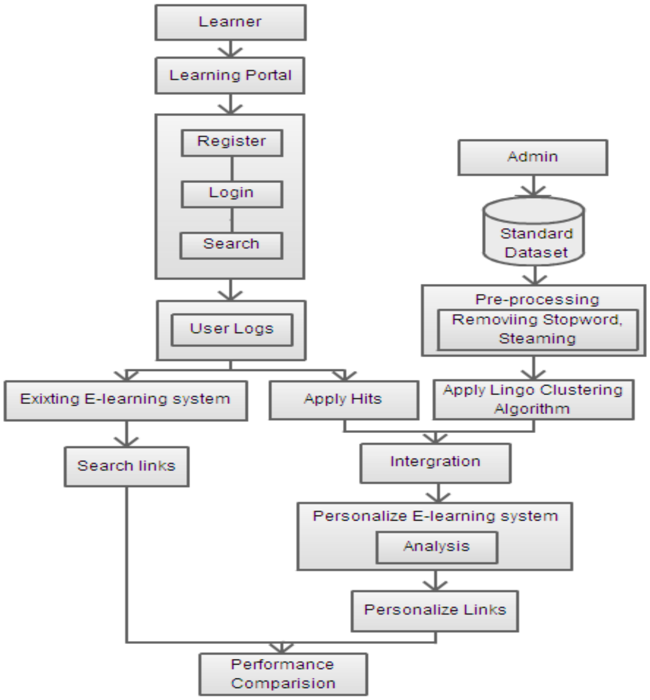


Fig. 3. Architecture of proposed approach

This combined approach gives a better performance as compared to existing E-learning system which is based on usage mining.

D. *Mathematical Model*

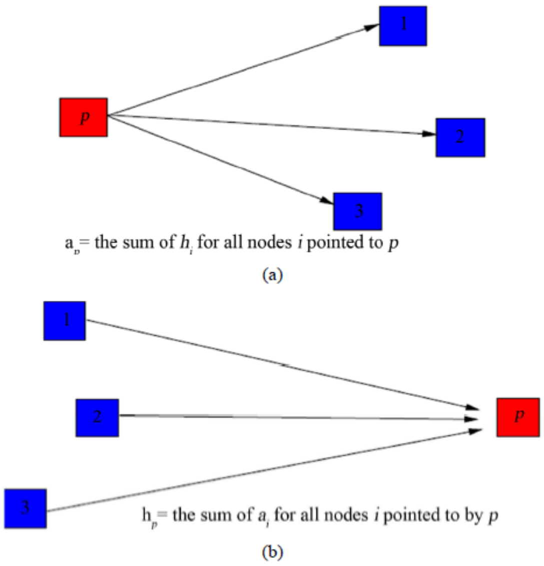
Algorithm: HITS (Hyper-link Induced Topic Search) algorithm
Input: Search String query.
Process:

1) *Sampling Step*

$$B(p) = \sum_{i=0}^{n} R(q)_i$$

where B(p) is the set of relevant pages and
R(q) is the result of given query.

2) *Iterative Step*



a_p = the sum of h_i for all nodes i pointed to p

(a)

h_p = the sum of a_i for all nodes i pointed to by p

(b)

Fig. 4. Dynamic weight updation

$For\ (each\ result\ of\ query\ )$

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

where H_q is Hub Score of a page

A_q is authority score of a page,

$I(p)$is set of reference pages of page p and

B(p) is set of referrer pages of page p,

end for

Output: max hit queried page.

Algorithm: Lingo Clustering Algorithm

Input:

User search query (q), term (t), word (w), Number of clusters k, $teaming\ words\ St_{SW}$, Set of stopwords

$$S_{SW} = \{"i", "a", "about", "an", "and", "as", ..\},$$

Process:

1) Preprocessing:

a) Remove stop words

If(q contains S_{SW}) then

Remove word.

b) $If(q\ contains\ St_{SW})\ then$

Rewrite word removing steaming

c) $If\ (q\ contains\ phrase)\ then$

Extract that phrase

2) Apply Lingo clustering Find clusters and label using Vector Space Model along with the Latent Semantic Indexing (LSI) technique.

a) Convert input query to $tf - idf$.

$$tf_{t,d} = number\ of\ ocuurance\ of\ term\ in\ document\ d$$

$$idf_t = log_{10} \frac{N}{dft}$$

where dft is document d that contain a term t

$$\mathrm{tf} - idf_{t,d} = \mathrm{tf}_{t,d} * idf_t$$

b) Create data structure indexed by document(d).

c) For each t

Update entry in score

$$score[d] = score[d] + tf = idf(t,d) * tf - idf(t,d)$$

Normalize score

$$Magnitude[d] = Magnitude[d] + tf - idf(t,d)\text{^}2$$

End for

d) For each d

Calculate cosine similarity

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Where,

q_i is the tf − idf weight of term i in the query.

d_i is the tf − idf weight of term i in the d.

end for.

Output: Clusters.

IV. Work Done

In this section we are discussing experimental analysis, performance metrics used and computed results etc.

A. Experimental Analysis:

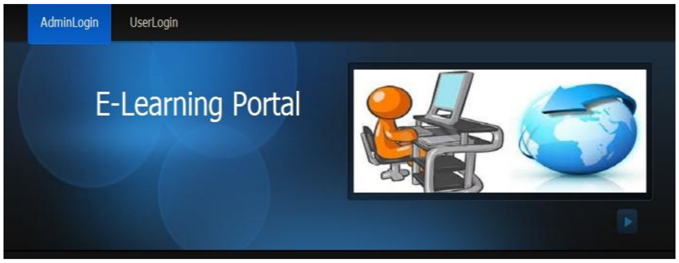
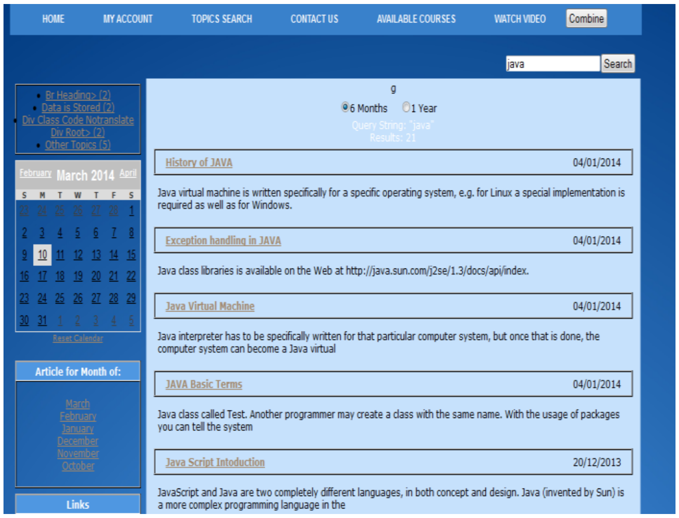


Fig. 5. Home page of Personalized E-learning system

As shown in Fig. 5, this is home page of personalized E-learning system. There are two different logins as user login and Admin login. After registration user can login and search for a particular topic



java

g

● 6 Months ○ 1 Year

Query String: "java"
Results: 21

History of JAVA 04/01/2014

Java virtual machine is written specifically for a specific operating system, e.g. for Linux a special implementation is required as well as for Windows.

Exception handling in JAVA 04/01/2014

Java class libraries is available on the Web at http://java.sun.com/j2se/1.3/docs/api/index.

Java Virtual Machine 04/01/2014

Java interpreter has to be specifically written for that particular computer system, but once that is done, the computer system can become a Java virtual

JAVA Basic Terms 04/01/2014

Java class called Test. Another programmer may create a class with the same name. With the usage of packages you can tell the system

Java Script Intoduction 20/12/2013

JavaScript and Java are two completely different languages, in both concept and design. Java (invented by Sun) is a more complex programming language in the

Fig. 6. Search results for java keyword with combined approach

As shown in Fig. 6, when users search for java keyword, he gets different links. When user selects a combined approach,

More accurate links are available to user, which results in personalized output.

B. *Results for precision of combined system against content based system:*

C. *Matrix measure*

To evaluate the effectiveness of the approach, performance is measured using two factors like Precision, Random Index.

Precision (P):

The percentage of retrieved documents that is in fact relevant to the query (i.e., "correct" responses).

$$precision = \frac{|\{Relevant\} \cap \{Retrived\}|}{|\{Retrived\}|}$$

Random Index (RI):

The Rand index measures the percentage of decisions that are accurate

$$RI = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

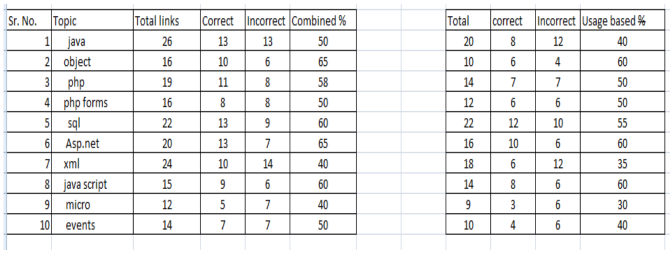
A true positive (TP) decision assigns two similar documents same cluster. True negative (TN) decision assigns two dissimilar documents to different clusters. (FP) decisions assign two dissimilar documents to same cluster. A (FN) decision assigns two similar documents to different clusters [7].

D. *Results of work done*

The comparative study between existing system based on usage mining and proposed system with combined approach are as shown in following figure.

Ambient (ambiguous entries) dataset is used as standard dataset for finding search results. Dataset contain many queries, results of few queries are as follows.

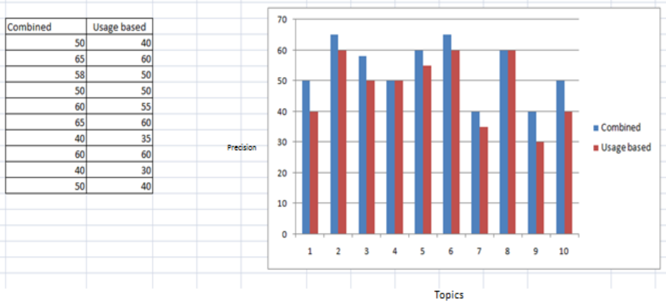
Results for precision of combined system against usage based system:



Sr. No.	Topic	Total links	Correct	Incorrect	Combined %	Total	correct	Incorrect	Usage based %
1	java	26	13	13	50	20	8	12	40
2	object	16	10	6	65	10	6	4	60
3	php	19	11	8	58	14	7	7	50
4	php forms	16	8	8	50	12	6	6	50
5	sql	22	13	9	60	22	12	10	55
6	Asp.net	20	13	7	65	16	10	6	60
7	xml	24	10	14	40	18	6	12	35
8	java script	15	9	6	60	14	8	6	60
9	micro	12	5	7	40	9	3	6	30
10	events	14	7	7	50	10	4	6	40

Fig. 7. Values for searched topics for precision

Graph for precision



Combined	Usage based
50	40
65	60
58	50
50	50
60	55
65	60
40	35
60	60
40	30
50	40

Fig. 8. Graph for precision analysis

V. Conclusion and Future Work

This paper proposes a combined approach for personalized E-learning system based on web usage mining and web content mining. In this approach weight associated with a hyperlink is assigned by considering equal percentage of web usage mining and web content mining. User logs are saved at server side, which helps to provide personalized output with area of interest of the user. HITS algorithm is applied to user logs to increase weight of that hyperlink. At the same time, contents of a web page are preprocessed and clustering results are combined with hits a result which gives more accurate results.

The performance of the system is evaluated under different settings and in comparison with the previous method which is based only on the usage mining. The application can be used for personalized recommendation to give personalized recommendation based on users browsing history. Throughout this paper we have discussed many aspects of research for personalized E-learning system. Based on existing limitations, in this paper a new web mining approach based on combination of web usage mining and web content mining is presented which is showing the better performance improvement as compared to the existing method.

In future we discover the learner's time distribution pattern to realize personalized curriculum organization, discover the learning behavior pattern to build up a series of feedback and motivation system and we will give different training according to different learners' levels.

References

[1] Yuewu Dong, Jiangtao Li, "Personalized Distance Education System Based on Web Mining", International Conference on Education and Information Technology (ICEIT 2010), IEEE, 2010

[2] Maurice D. Mulveny , Sarabjot S Anand and Alex G. Buchner, "Personalization on the Net using Web Mining," Communications of the ACM, vol. 43, No. 8, 2000.

[3] Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, vol. 3, No. I, February 2003.

[4] Saeko Nomura Satoshi Oyama Tetsuo Hayamizu Toru Ishida, “Analysis and Improvement of HITS Algorithm for DetectingWeb Communities”, Department of Social Informatics, Kyoto University Honmachi Yoshida Sakyo-ku, Kyoto, 606-8501 Japan@kuis.kyoto-u.ac.jp, ishida@i.kyoto-u.ac.jp

[5] R.Mahalakshmi, V.Lakshmi Praba, “A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO”, 6, August 2013

[6] Stanislaw Osi´nski and Dawid Weiss, “A Concept-Driven Algorithm for Clustering Search Results”, 2005 IEEE

[7] Sri Shilpa admanabhuni, Hima Bindu T, “Search Results Clustering: Comparison of Lingo and K-Means”, Volume 1, Issue 5, October 2013.

[8] K. Palanivel, S. Kuppuswami ,“ Service-Oriented Reference Architecture for Personalized E-learning Systems (SORAPES) ”, International Journal of Computer Applications (0975 – 8887) , Volume 24– No.5, June 2011

[9] A.Jebaraj Ratnakumar , “An Implementation of web personalization using web mining techniques”, Journal of Theoretical and Applied Information Technology, 2005 - 2010 JATIT.

[10] Mr.Ramesh Prajapati, “A Survey Paper on Hyperlink-Induced Topic Search (HITS) Algorithms for Web Mining”, International Journal of Engineering Research and Technology (IJERT), Vol. 1 Issue 2, April – 2012

[11] Kavita D. Satokar, Prof..S.z.Gawali, “Web Search Result Personalization using Web Mining”, International Journal of Computer Applications (0975 - 8887), Volume 2 - No.5, June 2010.

[12] K. Sridevi, R. Umarani, V.Selvi, “An Analysis of Web Document Clustering Algorithms”, International Journal of Science and Technology, Volume 1 No.6, December 2011.