EFFICIENT WEB DATA EXTRACTION USING CLUSTERING APPROACH IN WEB USAGE MINING

Neeraj Raheja[1] and V.K.Katiyar[2]

[1,] Associate Professor, [2] Professor
[1,2] Department of Computer Engineering, M.M.Engineering College, M.M.University, Mullana(Ambala), Haryana, India

Abstract: Web usage mining is used to record user behavior. These records are further used to extract data which helps in search engine optimization. In this research work we propose an approach in which web logs are used in cluster forms. These clusters are designed according to the user behavior records in web logs. Hence when we search from these clusters instead of complete web log, searching time gets reduced.

Keywords: Web Mining, Web usage mining Clustering, Weblog.

1. Introduction

1.1 Web mining: web mining which is a type of data mining is used to extract web data from web pages. As data mining basically deals with the structured form of data, web mining deals with the unstructured and semi-structured form of data. Web mining consist of three techniques i.e. web content mining, web structure mining and web usage mining for web data extraction[9][10] as shown in fig. 1

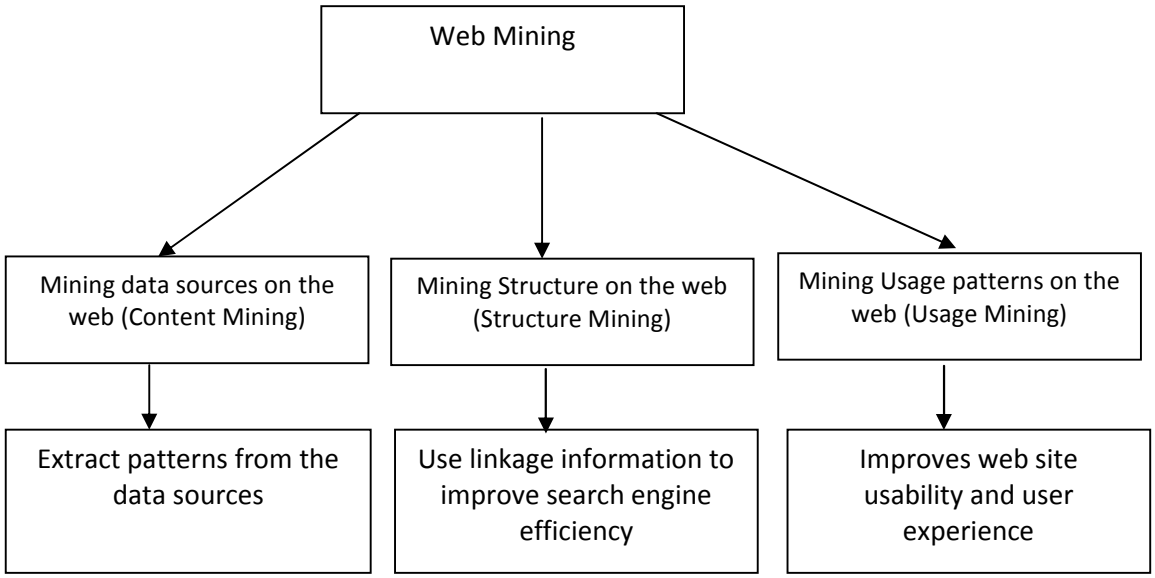


Figure 1: Web mining and its types

Web content mining deals with extraction of data from the content of WebPages based upon pattern matching. [9][10][12] Web structure mining deals with the linkage structure of the WebPages and used to extract information from these structures.[13] Finally web usage mining deals with user behavior i.e. record the user activities in web logs and these web logs are further used to extract important information.[9][10]

1.2 Web usage mining:

Web usage mining is the process to record the activities of the users while they are browsing and navigating through the Web. The basic aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [16] or to improve the Web structure and Web server performance. Examples of applications of such knowledge include improving designs of web sites, analysing system performance as well as network communications, understanding user reaction and motivation, and building adaptive Web sites.

The process of Web usage mining also consists of three main steps: (i) pre-processing, (ii) pattern discovery and (iii) pattern analysis.[9]

Weblog

Weblog is a file which is created according to the user visits or user activities on a webpage or website. A common weblog is shown in fig. 2, which consists of ip of visited website, date and time, access method, data or webpage accessed, web client or browser with version used and plateform used etc.[10][12].

1. 123.456.78.9 - [25/Apr/1998:03:04:41 -0500] "GET A.html HTTP/1.0" 200 3290 - Mozilla/3.04 (Win95, I)
2. 123.456.78.9 - [25/Apr/1998:03:05:34 -0500] "GET B.html HTTP/1.0" 200 2050 A.html Mozilla/3.04 (Win95, I)
3. 123.456.78.9 - [25/Apr/1998:03:05:39 -0500] "GET L.html HTTP/1.0" 200 4130 - Mozilla/3.04 (Win95, I)
4. 123.456.78.9 - [25/Apr/1998:03:06:02 -0500] "GET F.html HTTP/1.0" 200 5096 B.html Mozilla/3.04 (Win95, I)
5. 123.456.78.9 - [25/Apr/1998:03:06:58 -0500] "GET A.html HTTP/1.0" 200 3290 - Mozilla/3.01 (X11, I, IRIX6.2, IP22)

Figure 2: Web log structure

2. Related Work

Ida Mele[1] provides an approach for improving search-engine performance through static caching of search results, and helping users to find interesting web pages by recommending news articles and blog posts. A query covering approach was used to search the web pages from cache and web logs and searching time, recall and precision was calculated on behalf of that.

The author [2] proposes an indiscernibility approach in rough set theory to extract information from extended web logs to identify the origin of visits and the keywords used to visit a web site which will lead to better design of websites and search engine optimization.

The author [3] had done the work on data preprocessing in web usage mining. They presented a new algorithm called —USIA (User and Session Identification). It finds the user and session identification details. The same user is identified with the help of IP address and User ID. If the request is from the same IP address, then the algorithm concluded that the request is from same user. The session is identified based on the time in and time out period. This research work mainly focused on user identification for the particular session and series of web pages viewed by the user.

This author [4] focused on grouping the customer transactions by using the clustering technique. The set of transactions in a group has some similarities, so we can easily identified the customer behaviour and the web site analyst can able to understand the customer expectation and make the website customer friendly. In other point of view, make the website is more personalized and more user friendly. The researcher used the pattern based clustering approach to group the similar type of transactions.

The author [5] dealt with two types of groups one is Web Clustering Groups which groups the relative pages from the web server log files, the second is User Clustering Groups which groups the user who refers the same type of web pages. Divisive Hierarchical Clustering Algorithm is used to group the Web Log files and User of similar type. Then the association rule mining with support and confidence measure is applied to each group to fine the relationship among them.

This author [6] focused on the first phase of Web Usage Mining called Data Pre-processing and they suggested a novel approach for feature selection based on Rough set Theory for Web Usage Mining. The problem in web Log Files is their size and unwanted data. This paper used two algorithms Quick reduct and Variable Precision Rough Set Algorithm to identify the necessary data from the web log files, the actual process of feature selection. The k-means clustering algorithm is used to segment the similar patterns before applying the above two algorithms. So the algorithms are applied only to the group of similar items to identify the feature selection.

So, this technique given the optimal solution for eliminating the unwanted data in the web log files.

The author [7] mainly focused on the data pre processing step to remove the unnecessary data such as images, extra click events. Pattern discovery algorithms are used to eliminate the unwanted data from the web server log files. They taken the data from NASA website server log files and remove the unwanted data to improve the efficiency of the web log data analysing process. No specific data mining techniques are applied to web log files after pre processing. That work is open for future research workers.

The author [8] has done the comparative study on various sequential association rule mining algorithms with the various sequence and temporal constraints to predict the next request from the user. The result is affected based on the set of constraints. So, choosing the correct constraint given the better predictions result.

3. Proposed Architecture

This research work proposes an approach for web usage mining using cluster formulation. The results of cluster based web log searching are compared with the results of complete web log based searching i.e. caching of documents in the web log [1].

3.1 Complete web log searching algorithm

1. Read Input String (Si) as Keyword (Ki)
2. If (Si=NULL) Terminate / Halt. Re-Project Search Options
3. If (Si<>NULL) Establish Database Connection (DBCN)
4. if (ReturnType(DBCN)=NULL) DatabaseEngine Failed
5. If (DBCN)=Ri; {Ri=RecordSet)
6. Fetch / Retrieve RelatedRecord (RRi) from RLN(Relation).DBCN
7. Print RRi=>DataSetItem(i)
8. Move RecordLog(RLi) to ServerRepository(SR)
9. Compatability Check(CC) (Browser|Plugin|Add-On) => (True/False)
10. If (CC<>NULL) Print Results on WebClient(WC) {WC : Firefox/Chrome/IE/Safari/Opera}
11. Terminate with Success

3.2 Requirement for using clustering approach

Whenever a user searches some content or information according to a particular keyword, a lot of results are provided to user, some of these results are useful to user and most of them are not related. Also the normal habit of a user is to look for some top ranking results and ignore other. Same is the case when a complete web log is searched the results available will be more in numbers and searching time will also be more. Clustering approach partition the results available in the web log according to their ranking or popularity (i.e. number of times various users visit that webpage), hence it will search inside the top order cluster first and if sufficient number of results are not found, then it will go for the next one. In this way the searching time for most of the searches will be much less than the complete web log searching and most popular results will be available to the user.

3.2 Web log generation

1. A web log is created by the server according to user behavior or user visit in case of web usage mining.
2. The web log created is used to extract data according to popularity of data
3. Normally complete web log is searched to extract data or links, but in the proposed approach web log is searched in the form of clusters or partitions.

3.3 Steps used in proposed approach

1. When a user visits a web page, it is recorded in the web log as well as in a rank relevancy report. Rank relevancy report consists of record of all the referenced web pages along with their ranking.
2. The rank of the web page is calculated on the basis of referring the web page by various users (i.e. when any user visit a particular webpage the visit count of that web page gets increased).
3. The rank of the web page is used for selecting a particular webpage in a particular cluster.

3.4 Algorithm for the proposed approach

Consider n number of web pages is there

Phase 1: Relevancy rank report generation

```
I for i=1,i<=n,i++
  visit[i]=0
II if(k^th web page is visited by any user)
    then visit[k]=visit[k]+1;
III for i=1;i<=n;i++
      Sort (visit[i])
 IV for i=1;i<=n;i++
       Rank (i) = i
```

Phase 2: cluster formation

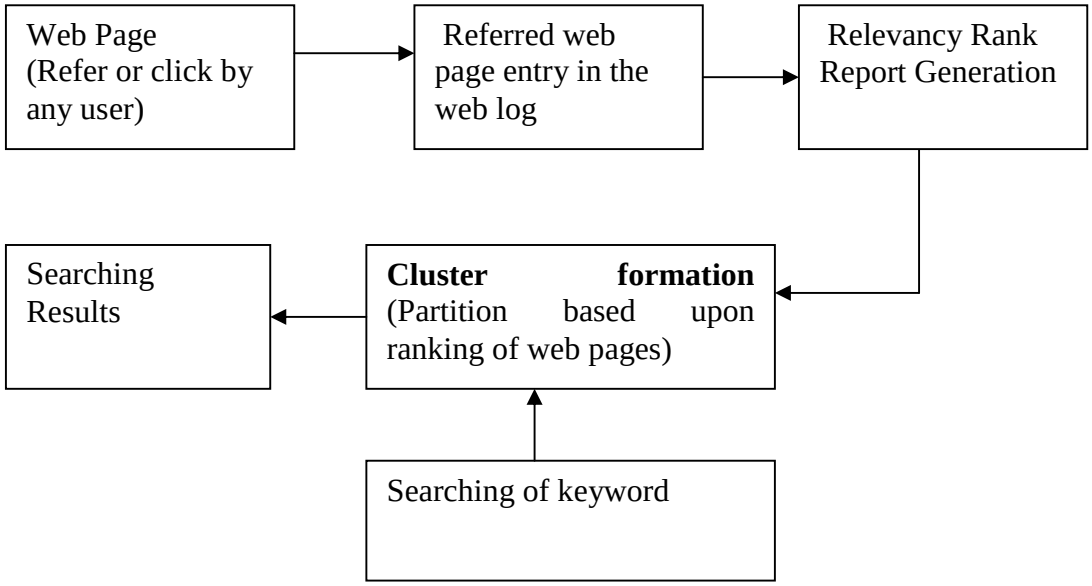
Consider c number of clusters is there and p number of pages is there in each cluster

V p=n/c

VI for i=1; i<=c;i++
for (j=(p*(i-1))+1;j<=p*i;j++)
(Cluster[i], rank[j])
(i.e cluster[i] consist of p number of pages according to rank)
VII if (k^{th} page is visited by the user) goto step II

consist of rank of web page on the basis of number of times the web page is visited by the user (fig 4). Then the result of both existing (complete web log searching time) as proposed in [1] is shown in fig. 5 and proposed (cluster based, in which 4 clusters are created) are performed as shown in fig 6. Finally in fig. 7 comparison of existing[1] and proposed approaches is shown

3.5 Flowchart for proposed approach



3.6 Searching time calculation

For showing the results the searching time was calculated on basis of query or keyword entered by the user and till the results are obtained from the web log i.e. database generated.

4 Experimental Results and Discussions

For showing the results of the proposed approach a website of 15 web pages is created in php. A Search engine is developed to get input or keyword from the user (Fig 3). A relevancy rank report is generated on the basis of referring of these web pages which

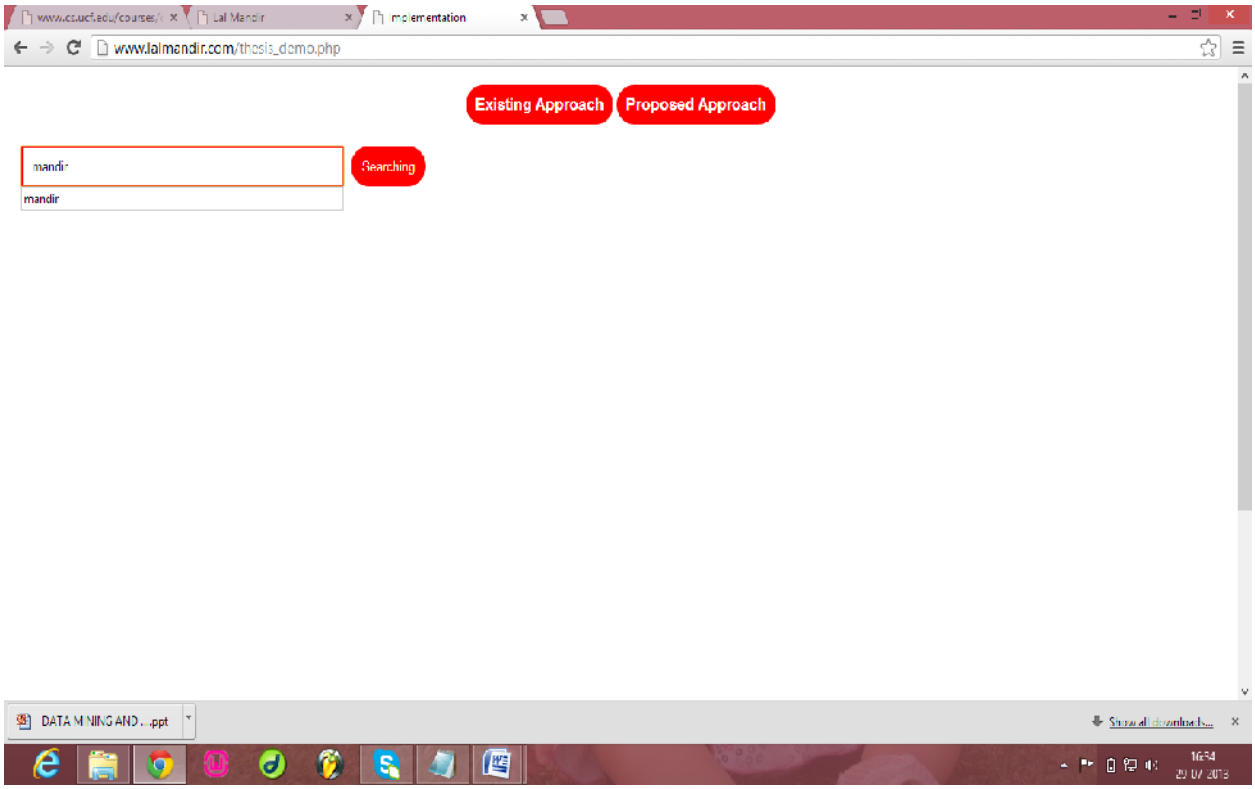


Fig. 3 : Search engine for user input or keyword

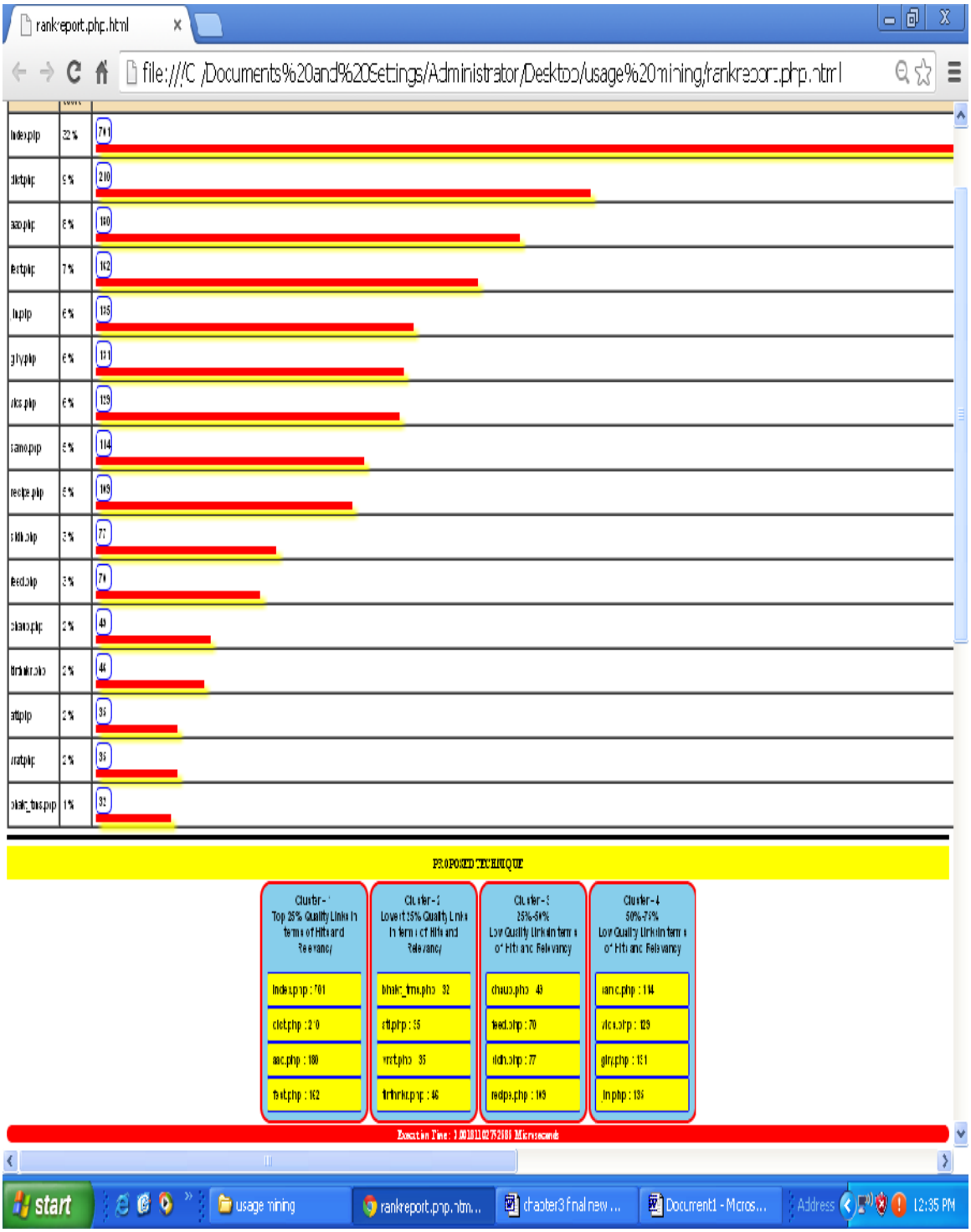


Fig 4: Relevancy rank report

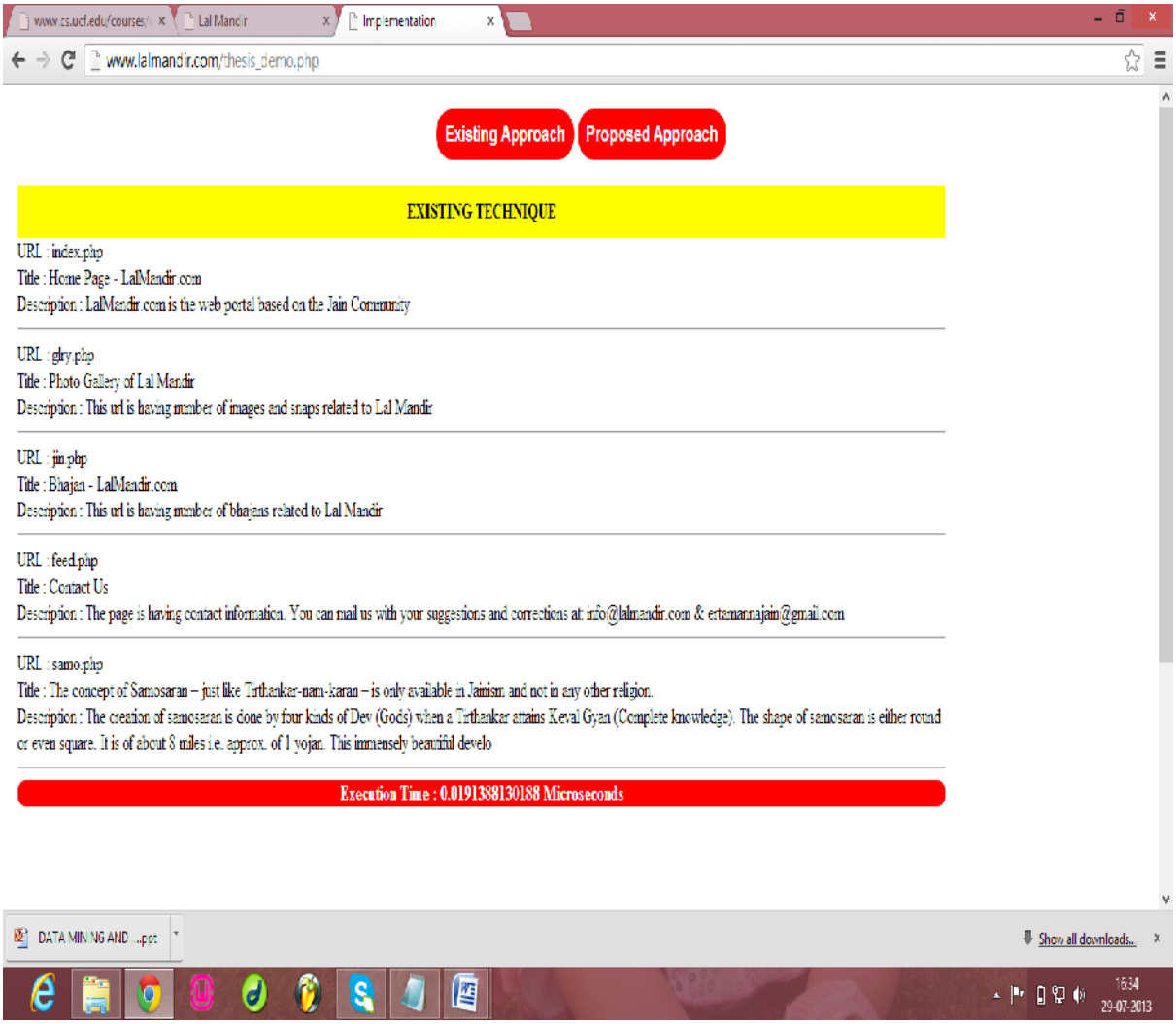


Fig 5: Searching of keyword from complete web log

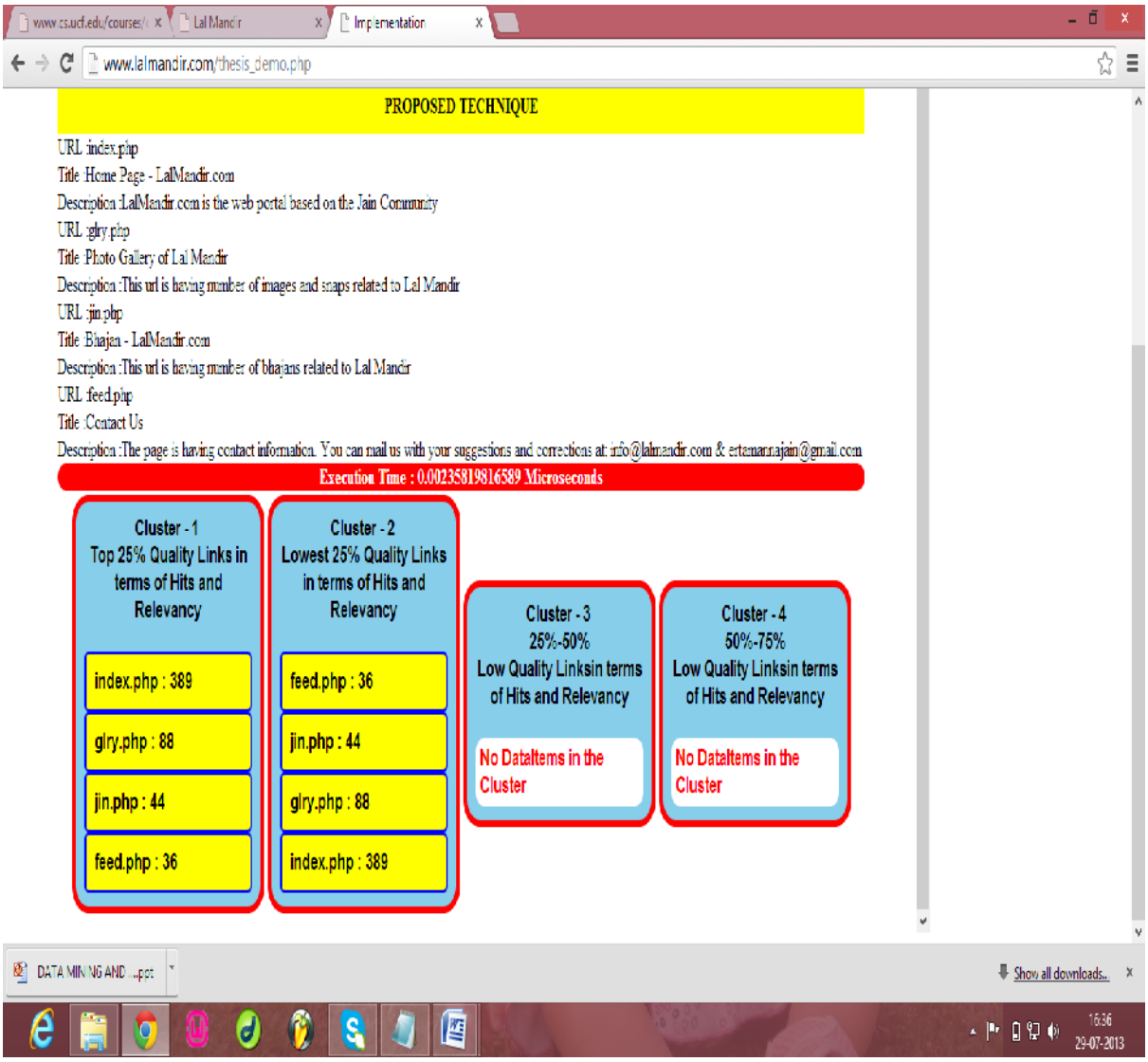


Fig 6: Searching of keyword from proposed approach

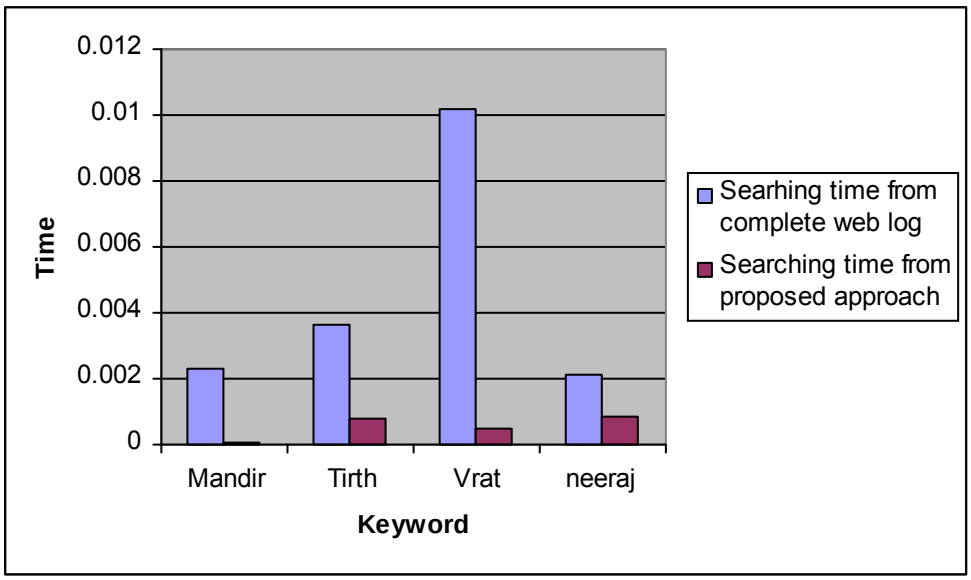


Fig 7: Comparison of complete web log searching and cluster searching time(ms)

Depending upon the rank relevancy report, more wonderful results may be obtained like checking the popularity of a web page or web site in a particular time slot which will provide current popular data and clusters will be formed on the basis of that.

5 Conclusion and future scope

This research work proposes an approach for web usage mining based upon web log partition. It takes less time and provides popular results in accordance with the existing approach. Some more results may be obtained if the number of cluster formed are changed i.e. from 4 clusters formed in our approach can be changed to 6, 8 or more. However recall and precision may be affected by changing the number of clusters i.e. either may be improved or decayed.

References

[1] Ida Mele, Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content, ACM, WSDM'13, pp. 765-769 Rome, Italy, February 2013.

[2] Jeeva Jose and P. Sojan Lal(2013) Extracting Extended Web Logs to Identify the Origin of Visits and Search Keywords , Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, pp 435-441.

[3] Zhang Huiying, Liang Wei.An (2004). Intelligent Algorithm of Data Pre-processing in Web Usage Mining. In Proceeding of the 5th World Congress on Intelligent Control and Automation. pp. 15-19. Hangzhou, P.R. China.

[4] Yinghui Yang and Balaji Padmanabhan. (2005). GHIC: A Hierarchical Pattern-Based Clustering Algorithm for Grouping Web Transactions. IEEE Transactions on Knowledge and Data Engineering, Vol 17, No. 9.

[5] Yi Dong, Huiying Zhang and Linnan Jiao. (2006). Research on Application of User Navigation Pattern Mining Recommentation. In Proceeding. of the 6th World Cogress on Intelligent Control and Automation. Dalian, China.

[6] Hannah Inbarani H., Thangavel K., and Pethalakshmi A. (2007). Rough Set based Feature Selection for Web Usage Mining. International Conference on Computational Intelligence and Multimedia Applications.

[7] Suneetha K. R., and Krishnamoorthi R. (2009). Identifying User Behavior by Analysizing Web Server Access Log File. IJCSNS International Journal of Computer Science and Network Security, Vol 9, No.4.

[8] Wang Yong Li and Zhanhuai Zhang Yang. (2005). Mining Sequential Association-Rule for Improving WEB Document Prediction. In Proceedings of the Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05).

[9] Kosala R., Blockeel H., (2000). Web mining research: a survey. SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2(1), pp. 1–15.

[10] J.Srivatsava, R.Cooley, M.Deshpande and P.N. Tan,(2000) "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." ACM SIGKDD Explorat. Newsletter,pp. 12-23.

[11] Web Data Extraction, Applications and Techniques: A Survey by Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner published at ACM Computing Surveys, Jul 2012.

[12] L.K. Joshila Grace, V.Maheswari and Dhinaharan Nagamalai(2011) “Analysis of web logs and web user in web mining”, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1.

[13] Yuefeng Li and Ning Zhong: Web Mining Model and Its Applications for Information Gathering, Knowledge-Based Systems 17, pp. 207–217, 2004.

[14] Rekha Jain and Dr. G. N. Purohit,”Page Ranking Algorithms for Web Mining, International Journal of Computer Applications”,ISSN: 0975 – 8887, Volume 13– No.5, pp. 22–25, January 2011.

[15] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, “Mining access patterns efficiently from web logs,” in PADKK ’00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, 2000, pp. 396-407.