International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 3 Issue 4 July 2013
Manuscript ID : 2249054XV3I4072013-01

A CONTEMPORARY STUDY ON CLUSTERING IN WEB USAGE MINING

Mr. Neeraj Raheja

CSE Deptt.

Maharishi Markandeshwar University, Mullana

Haryana, India

Tamanna Jain

M. Tech. (CSE) Student

Maharishi Markandeshwar University, Mullana, Haryana, India

ertamannajain@gmail.com

Abstract

The distension of the World Wide Web (Web for short) has accrued in a large amount of data that is now in general freely feasible for user access. The various types of data have to be managed and catalogued in such a way that they can be annexed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be altered such that they better suit the demands of the Web. This manuscript explains the concept of web usage mining and the role of clustering in mining to make the search results better as compared to the traditional technique of mining.

Keywords: Clustering, Usage Mining, Web Mining.

1. INTRODUCTION

The wide approbation of the Internet has fundamentally modified the ways in which we communicate, collect information, conduct businesses and make purchases. As the use of the World Wide Web and email skyrocketed, computer scientists and physicists rushed to differentiate this new phenomenon. While initially they were startled by the tremendous variety the Internet manifested in the size of its features, they soon discovered a widespread pattern in their measurements: there are many small elements that are contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others.

2. WEB USAGE MINING

Web Mining is based on knowledge discovery from web. It is extract the knowledge framework represents in a proper way. Web mining is like a graph and all pages are node & each connects with hyperlinks. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. By using web mining easily extract all features and information about multimedia before this web mining difficult to extract information in proper way from web. We search the any topic from web difficult to get accurate topic information but Now’s day it is easy to get the proper information about any things.

Web mining can be categorized in to three area of interest based on which part of the web to mine:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

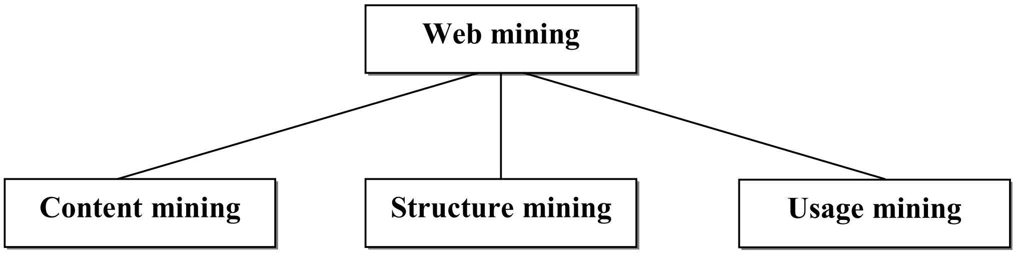


Figure 1.1: Categories of Web Mining

Web usage mining is an important technology for understanding user's behaviours on the web. Obtained user access patterns can be used in variety of applications, for example, one can keep track of previously accessed pages of a user. These pages can be used to identify the typical behaviour of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining. Mass customization and personalization performed by dynamic Content Web site by making clusters of users with similar access patterns and by adding navigational links.

Frequent access behaviour for the users can be used to identify needed links to improve the overall performance of future accesses. Pre-fetching and caching policies can be made on the basis of frequently accessed pages to improve latency time. In addition to modifications to the linkage structure, common access behaviours of the users can be used to improve the actual design of web pages and for making other modifications to a Web site. Moreover, usage patterns can be used for business intelligence in order to improve sales and advertisement.

It is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of inter related files on one or more web servers. It automatically generates the data stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content & site structure.

Web usage mining aims at utilizing data mining techniques to discover the usage patterns from web based application. The technique is to predict the user behaviour when he interacts with the web. It has three phases. They are:

- Pre-processing
- Pattern Discovery

- Pattern Analysis

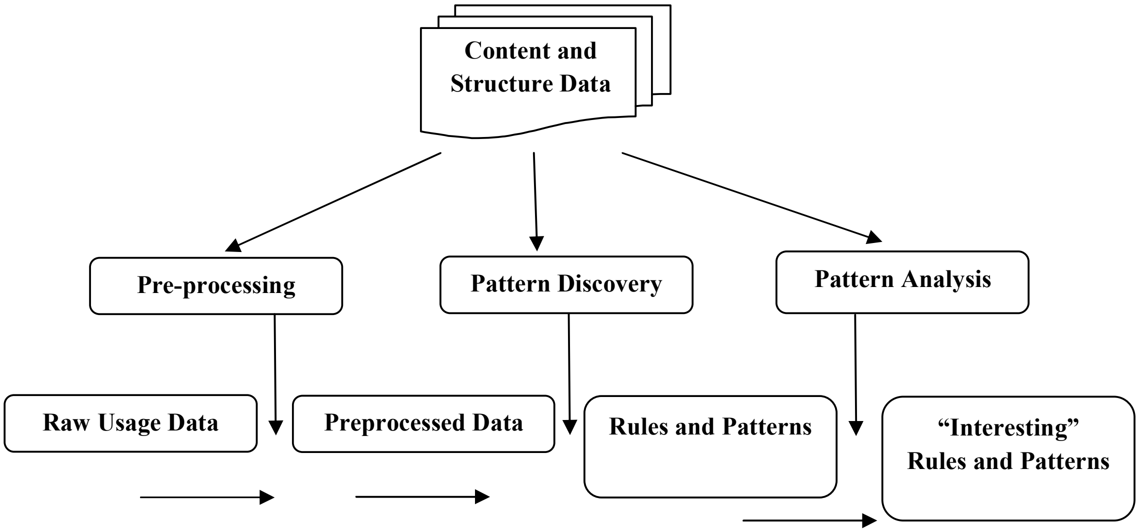


Figure 2.1: Web Usage Mining Process

3. WHY WEB USAGE MINING?

In this paper, we will emphasize on Web usage mining. Reasons are very simple: With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business. The growth of some E-businesses is astonishing, considering how E-commerce has made Amazon.com become the so-called "on-line Wal-Mart". Unfortunately, to most companies, web is nothing more than a place where transactions take place. They did not realize that as millions of visitors interact daily with Web sites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

Usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level.

Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. This enables an in-depth log to complete analysis of a company's productivity flow. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.

4. HOW TO PERFORM WEB USAGE MINING

Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data (as discussed below). Web server log files were used initially by the webmasters and system administrators for the purposes of "how much traffic they are getting, how many requests fail, and what kind of errors are being generated", etc. However, Web server log files can also record and trace the visitors' on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as "from what search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors?"

Web log file is one way to collect Web traffic data. The other way is to "sniff" TCP/IP packets as they cross the network, and to "plug in" to each Web server.

After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining techniques such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted.

5. CLUSTERING

Clustering is a well-studied data mining problem that has found applications in many areas. For example, clustering can be applied to a document collection to reveal which documents are about the same topic. The objective in any clustering application is to minimize the inter-cluster similarities and maximize the intra-cluster similarities. There are different clustering algorithms each of which may or may not be suited to a particular application.

Clustering identifies visitors who share common characteristics. After you get the customers'/visitors' profiles, you can specify how many clusters to identify within a group of profiles, and then try to find the set of clusters that best represents the most profiles.

Besides information from Web log files, customer profiles often need to be obtained from an on-line survey form when the transaction occurs. For example, you may be asked to answer the questions like age, gender, email account, mailing address, hobbies, etc. Those data will be stored in the company's customer profile database, and will be used for future data mining purpose. An example of clustering could be:

- 50% of clients who applied discover platinum card in /discovercard/customerService/newcard, were in the 25-30 age group, with annual income between $40,000 – 50,000.

Clustering of client information can be used on the development and execution of future marketing strategies, online and/or off-line, such as automated mailing campaign.

In clustering, there is no externally defined notion of correctness. There are a huge number of ways in which a training set could be partitioned. Some of these are better than others.

The classical methods suggest members of a cluster should resemble each other more than resemble members of other classes.

Hence a good partition should

- Maximise similarity within classes
- Minimise similarity between classes.

The traditional clustering paradigm pertains to a single dataset. Recently, attention has been drawn to the problem of clustering multiple heterogeneous datasets where the datasets are related but may contain information about different types of objects and the attributes of the objects in the datasets may differ significantly. A clustering based on related but different object sets may reveal significant information that cannot be obtained by clustering a single dataset.

6. PROPOSED ALGORITHM

In proposed algorithm, suppose there are n tuples. A fitness value is assigned to each tuple using the fitness function. Based upon this fitness value the tuples will be assigned to the clusters. If the fitness value [31] of the tuple is equal to or nearly equal to the threshold value of the generated set of random clusters then only the tuple will be assigned to the cluster otherwise tuple is assigned to the outlier cluster. If there are many clusters in the outlier cluster then a similarity is calculated among theses clusters and outlier is detected. In this approach, tie can also occur i.e. if a tuple belongs to two clusters then we can arbitrarily assign this tuple to any one cluster.

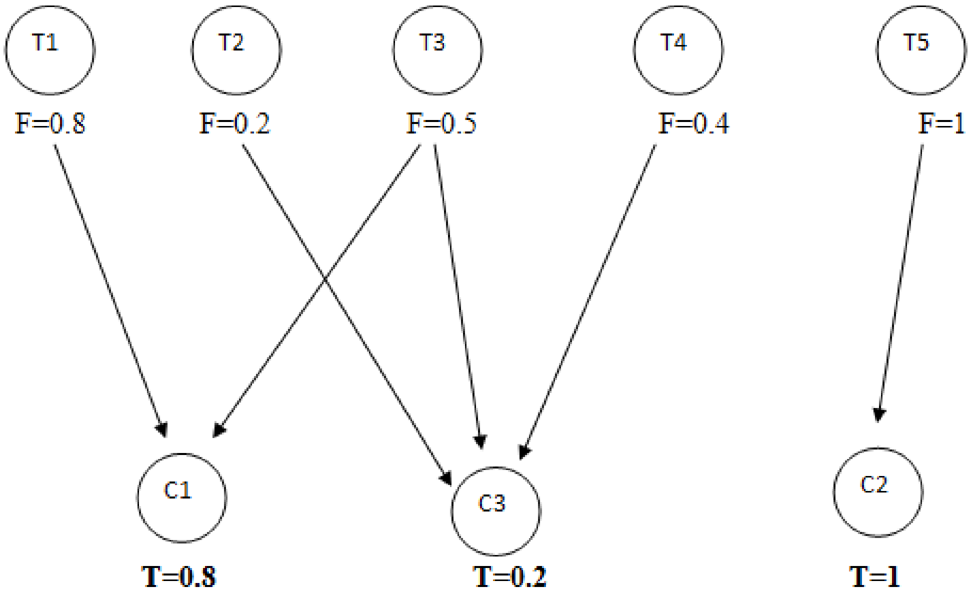


Fig 6.1 Proposed Approach

T= Threshold

F= Fitness Value

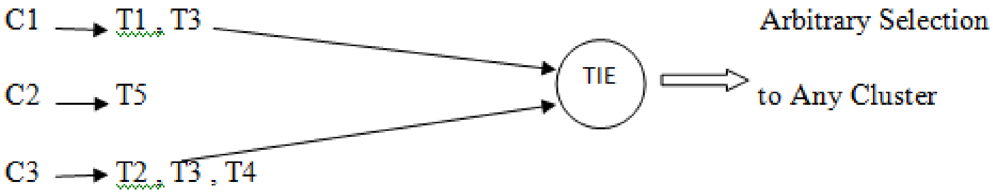


Figure 6.2: Analysis of Fitness and Selection

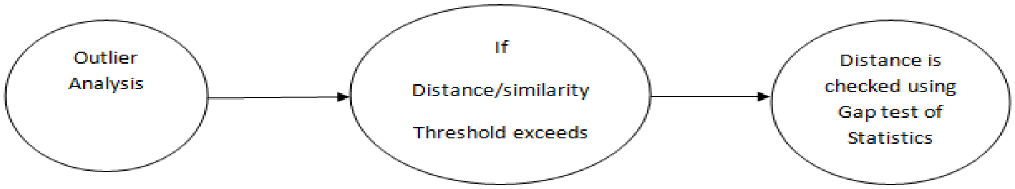
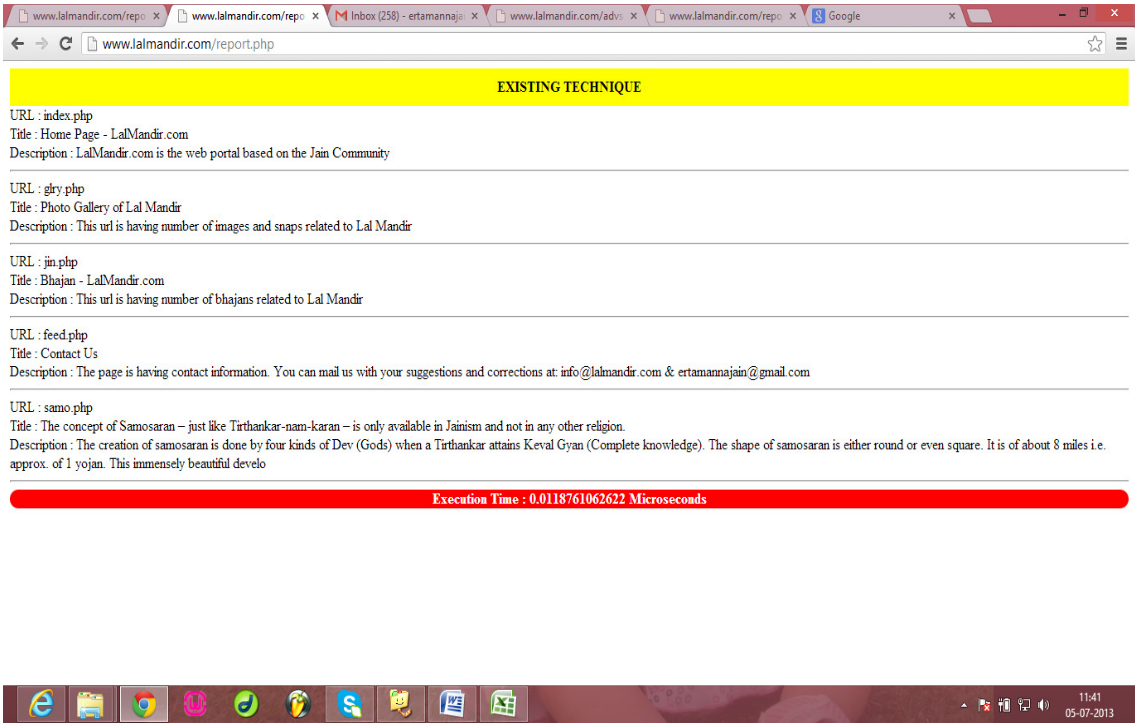


Figure 6.3: Outlier Detection

The training dataset is selected either randomly or sequentially from the data warehouse. Then calculate fitness value of each tuple. Further clusters and outliers are detected with the help of filtering module. Here filtering module is the similarity measure function. With the help of similarity measure comparison of cluster and tuple will take place. After the generation of clusters and outlier detection final statistics and report is generated.

7. RESULTS

The proposed algorithm presented in this manuscript is applied on a religious web portal www.lalmandir.com. The results are then compared on the basis of time metrics. The time taken to make the searching in existing technique is more as compared to the proposed technique. The snapshots of the two is shown in fig 7.1 and fig 7.2



EXISTING TECHNIQUE

URL : index.php
Title : Home Page - LalMandir.com
Description : LalMandir.com is the web portal based on the Jain Community

URL : glry.php
Title : Photo Gallery of Lal Mandir
Description : This url is having number of images and snaps related to Lal Mandir

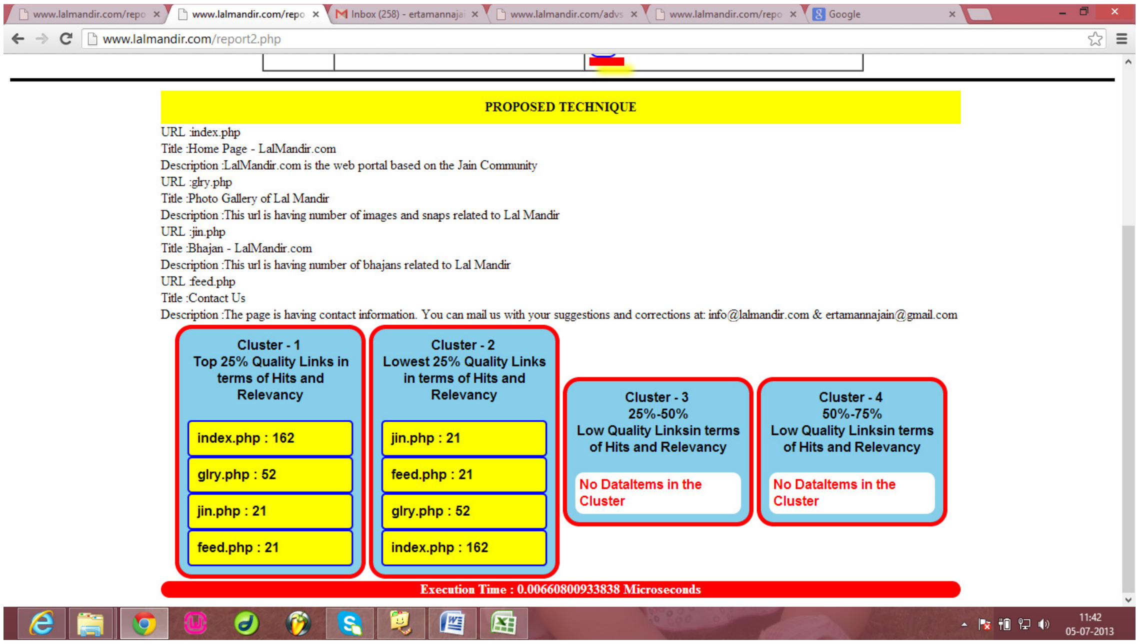
URL : jin.php
Title : Bhajan - LalMandir.com
Description : This url is having number of bhajans related to Lal Mandir

URL : feed.php
Title : Contact Us
Description : The page is having contact information. You can mail us with your suggestions and corrections at: info@lalmandir.com & ertamannajain@gmail.com

URL : samo.php
Title : The concept of Samosaran – just like Tirthankar-nam-karan – is only available in Jainism and not in any other religion.
Description : The creation of samosaran is done by four kinds of Dev (Gods) when a Tirthankar attains Keval Gyan (Complete knowledge). The shape of samosaran is either round or even square. It is of about 8 miles i.e. approx. of 1 yojan. This immensely beautiful develo

Execution Time : 0.0118761062622 Microseconds

Fig 7.1 Existing approach



PROPOSED TECHNIQUE

URL :index.php
Title :Home Page - LalMandir.com
Description :LalMandir.com is the web portal based on the Jain Community
URL :glry.php
Title :Photo Gallery of Lal Mandir
Description :This url is having number of images and snaps related to Lal Mandir
URL :jin.php
Title :Bhajan - LalMandir.com
Description :This url is having number of bhajans related to Lal Mandir
URL :feed.php
Title :Contact Us
Description :The page is having contact information. You can mail us with your suggestions and corrections at: info@lalmandir.com & ertamannajain@gmail.com

Cluster - 1 Top 25% Quality Links in terms of Hits and Relevancy	Cluster - 2 Lowest 25% Quality Links in terms of Hits and Relevancy	Cluster - 3 25%-50% Low Quality Linksin terms of Hits and Relevancy	Cluster - 4 50%-75% Low Quality Linksin terms of Hits and Relevancy
index.php : 162	jin.php : 21	No DataItems in the Cluster	No DataItems in the Cluster
glry.php : 52	feed.php : 21		
jin.php : 21	glry.php : 52		
feed.php : 21	index.php : 162		

Execution Time : 0.00660800933838 Microseconds

Fig 7.2 Proposed approach

8. CONCLUSION

In the present manuscript we have analyzed the performance of the algorithm based on the time metrics. The performance analysis of the algorithm is shown in fig 8.1. A comparison between the traditional approach and the existing approach is done shown in fig 8.2 to judge the performance of the two.

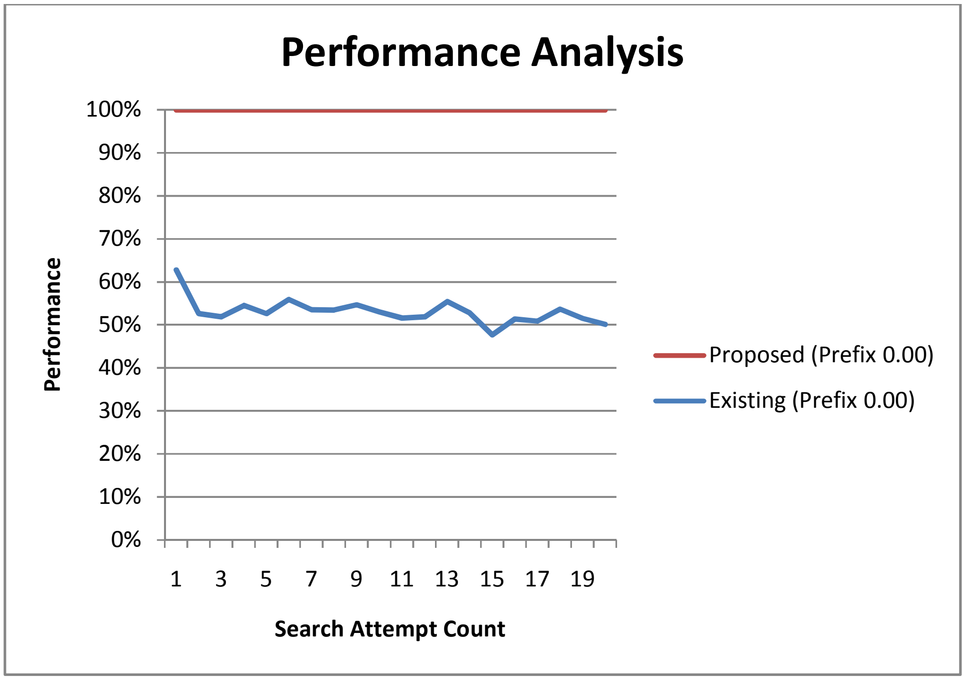


Fig 8.1 performance analysis

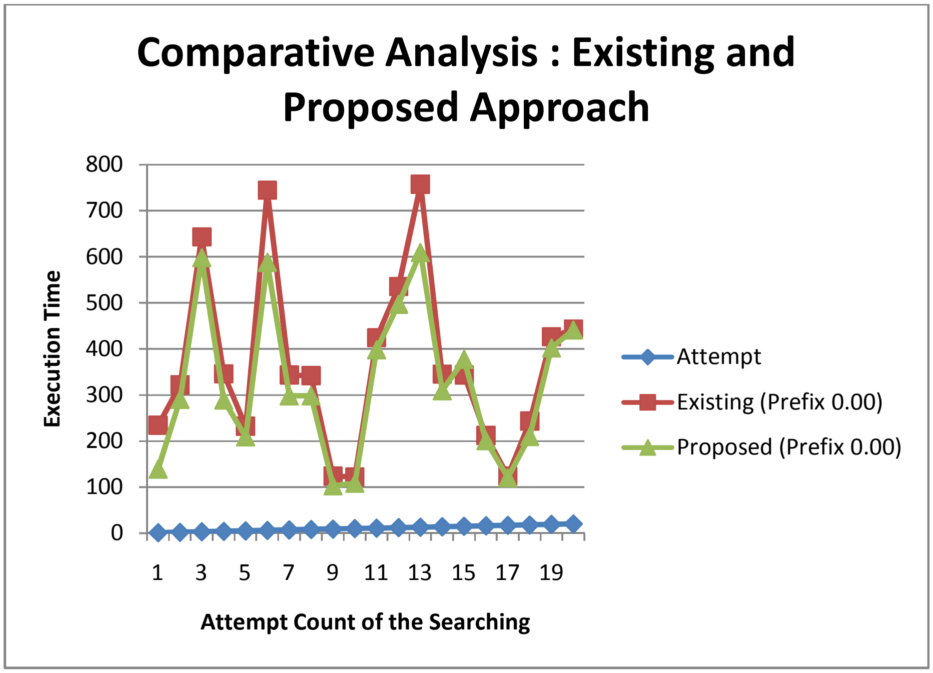


Fig 8.2 comparative analysis

REFERENCES

[1] http://www.google.co.in/url?sa=t&rct=j&q=why%20web%20usage%20mining&source=web&cd=9&cad=rja&sqi=2&ved=0CGMQFjAI&url=http%3A%2F%2Fwww.ef.uns.ac.rs%2Fmis%2Farchive-pdf%2F2010%2520-%2520No1%2FMIS2010_1_5.pdf&ei=LjneUYXvAsi6kQWjlYDgCQ&usg=AFQjCNGsOUjZazU_qptKqAcfnw-mf_RQ0Q&bvm=bv.48705608,d.dGI

[2] http://www.google.co.in/url?sa=t&rct=j&q=why%20web%20usage%20mining&source=web&cd=4&cad=rja&sqi=2&ved=0CEEQFjAD&url=http%3A%2F%2Fmmlabold.ceid.upatras.gr%2Fcourses%2FAIS_SITE%2Ffiles%2F3%255CWeb%2520Usage%2520Mining%2520as%2520a%2520Tool%2520for%2520Personalization-%2520A%2520Survey.pdf&ei=LjneUYXvAsi6kQWjlYDgCQ&usg=AFQjCNFEXLX_tcHei19y0m3LO377wqfq1A&bvm=bv.48705608,d.dGI

[3] http://www.google.co.in/url?sa=t&rct=j&q=why%20web%20usage%20mining&source=web&cd=8&cad=rja&sqi=2&ved=0CFoQFjAH&url=http%3A%2F%2Fwww.rimtengg.com%2Fiscet%2Fproceedings%2Fpdfs%2Fdatabase%2F73.pdf&ei=LjneUYXvAsi6kQWjlYDgCQ&usg=AFQjCNERupfvscS7VjlNyvD_fCJkhnBU_Q&bvm=bv.48705608,d.dGI

[4] http://www.web-datamining.net/usage/

[5] http://www.sciencedirect.com/science/article/pii/S0169023X08000104

[6] http://en.wikipedia.org/wiki/Web_mining