MAPTICS: A Combined Approach of Association Rule & Clustering for Large Dataset in Web Mining

Shukla Devashree Krishnakumar
Student of Master of Engineering
Dept. of IT System & Network Security
SVIT – Vasad , Gujarat , India

Ms. Sneha Gaywala
Assistant Professor
Dept. of IT System & Network Security
SVIT – Vasad , Gujarat , India

***Abstract*— There is currently a large amount of useful data available in WWW and it is also increasing day by day due to its friendliness to the people. So it is required to handle those store of data quickly, effectively, correctly on time. Web mining is the process of applying data mining techniques to extract useful information from web logs. Web usage mining refers to analyzing web log data to identify user behaviour, frequent patterns which will solve many web related problems, There are large number of data mining techniques developed & refined with passage of time. Apriori is very basic algorithm based on association rule mining. It has much variant of it. This paper aims to optimize the performance of one of its variant using a density based clustering technique. The proposed algorithm will reduce memory requirement, improve execution time and accuracy.**

Keywords*— Web mining, cluster algorithm, Modified Apriori, frequent pattern mining

I. INTRODUCTION

As shown in figure 1 , *Data mining* is evaluated over time. In beginning 1995, data mining restricted to Relational databases (RDB) only. After that, new inventions are made in Information technology , in 2000 data became rich like multimedia files, images, videos etc. Then that was the start up of *web mining*. After 2005 , there was great increase in internet users which provided users scalability, mobility, reliability, friendliness so new direction to internet is found , that was called *Semantic web*. The Semantic Web is a collaborative movement led by international standards body the World Wide Web Consortium (W3C).Implementation of semantic web lead to a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Semantic web is gift of Tim Berners Lee - inventor of world wide web.

Data Mining : is the process of extracting patterns (knowledge) from data. Contains four types of techniques to extract data. Association rule based , Classification, Clustering, Regression.

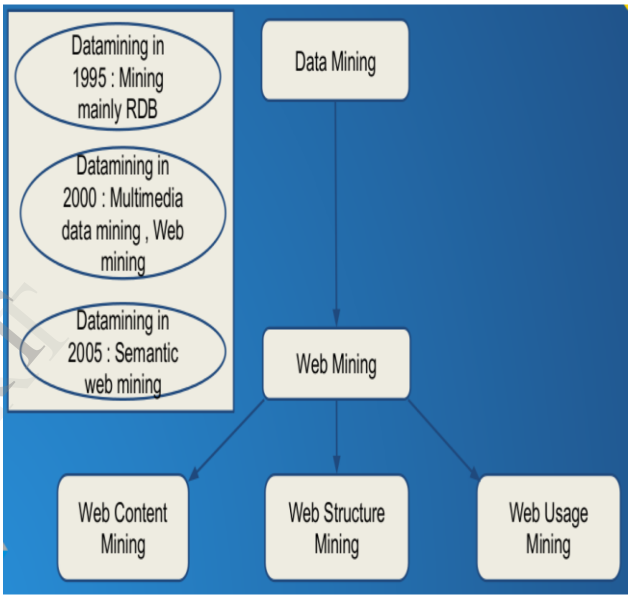

Datamining in 1995 : Mining mainly RDB
Datamining in 2000 : Multimedia data mining , Web mining
Datamining in 2005 : Semantic web mining
Data Mining
Web Mining
Web Content Mining
Web Structure Mining
Web Usage Mining

Figure 1 : Evaluation of data mining over time

Web Mining : is the application of data mining techniques to discover patterns from the Web data / web logs like web server log , web user/usage log , proxy server log etc.

Semantic Web Mining : mining the semantic web & intelligent web mining, mining web pages / data sources to develop effective web & to better understand the information. From literature survey of mining techniques i can conclude that web mining is an application of data mining techniques & implements Semantic web. From the last decade there was countless researches are made to improve the performance of internet on WWW.

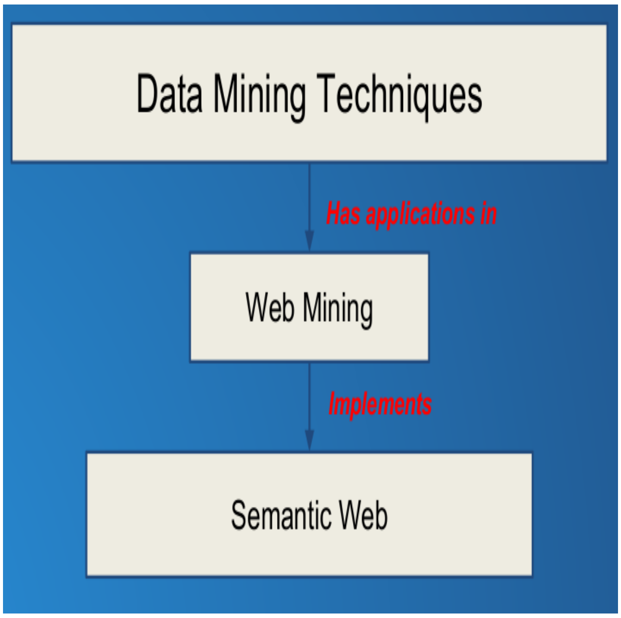


Figure 2 : Summary of current web state

II. WEB MINNING

1) Web mining : Used to extract web data from web pages. The objective & purpose of web mining gives classification as follows.

a) Web Content mining : Also called Text mining. Content mining is the process of mining , scanning, analyzing of texts, images , graphs of a web page to determine its significance to user. Natural language processing & Information retrieval are widely used technology for web content mining.

b). Web Structure mining: Used to recognize that how the web pages are connected. It is process of extracting information from links of web pages i.e. Hyperlinks. It is the process of using graph theory to analyze the node and connection structure of a web site. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (eXtensible Markup Language) tags within the web page[1].

c). Web Usage mining : Also called web log mining, aim is to discover interesting and frequent user access patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions. It is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web. The usage data records the user's behavior when the user browses or makes transactions on the web site. Involves the automatic discovery of patterns from one or more Web servers.

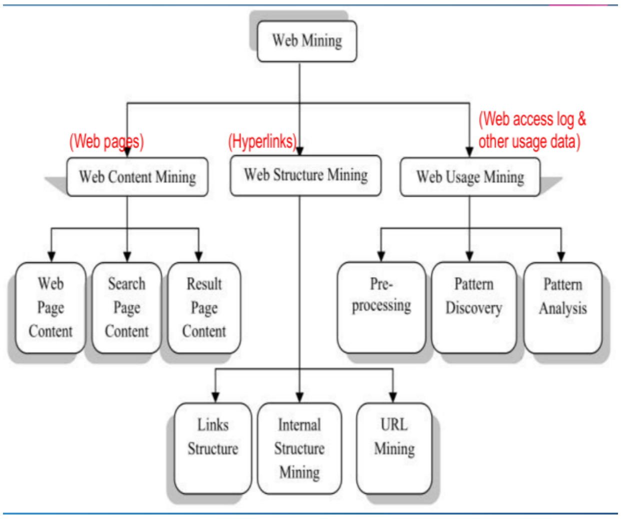


Figure 3 : Classification of web mining

2) Web log files : Files that contain information of website visitor activity. Log files are created by web servers automatically. Log entries are stored as a line of text in log file. Log file size will be 1KB to 100MB.

a) Location of weblog file: 3 different locations are available.

i) *Web server logs:* Web log files provide most accurate and complete usage of data to web server. The log file do not record cached pages visited, kept closed by web server.

ii) *Web proxy server:* It takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Establishment of Proxy server is a difficult process.

iii) *Client browser:* Log file resides in client's browser window. HTTP cookies are information generated by a web server & stored in computer for future access.

b) *Type of web log file :* There are four types of server logs.

i) *Access log file:* Data of all incoming request and information about client of server. Access log records all requests that are processed by server.

ii) *Error log file:* list of internal error.

iii) *Agent log file:* Contains Information about user's browser, browser version.

iv) *Referrer log file:* Provides information about link and redirects visitor to site.

c) Web log file format : Web log files are plain text file which record information about each user. Display of log files data in three different formats.

i) W3C Extended log file format
ii) NCSA common log file format
iii) IIS log file format

III. MODIFIED APRIORI - ASSOCIATION RULE BASED ALGORITHM

Traditional Apriori algorithm used widely by researchers & has some limitations found for large data set mining. To overcome the problem of increased time & memory requirement for large data set ,many algorithms are developed. one of them is developed by of the paper which is called Modified Apriori. The author put restriction on transaction made in database. Pseudo code is as below.

```
Variables:
C k : Candidate item-set of size k
L k : frequent item-set of size k
L 1 = {frequent items};

Process:
For (k = 1; L k !=∅; k++) do begin
C k+1 = candidates generated from L k ;
For each transaction t in database do
If ( t== input set) then
{
Increment the count of all candidates in C k+1
}
Those are contained in t
L k+1 = candidates in C k+1 with min_support
End
Return ∪ k L k ;
```

Table 1 : Pseudo code of Modified Apriori algorithm

The performance of algorithms are examined using N cross Validation method and in context of Accuracy , Memory usage , Build time & Search time.

a) Accuracy is calculated using the total number of correctly classified objects versus the total sample produced to classify.

b) The total memory resources required to execute the given algorithm is known as Memory usage of the system. Which is defined as the peak memory required executing the system or algorithm is known as memory uses.

c) Total time required to develop data model using the input data is known as model build time. That is estimated using the elapse time between initialization of algorithm and finishing the model building.

d) Search time or prediction time is defined as the time required predicting a value after accepting some parameters.

The summary of results are as follows.

a) Found that when the size of data increases the developed model perform poor results and some are provide more accurate results.

b) In Frequent set mining our modify Apriori algorithm takes less time than Apriori algorithm for search patterns and building data models.

c) Memory used is directly proportional to the size of data.

d) Accuracy of results are not assured in every case.

IV. PROBLEM STATEMENT

It can be notice that whether the modified algorithm has improved the performance of original Apriori technique , there is also scope of improvement & optimize the performance. The algorithm gives weaker performance when the data set size increases. The algorithm can also reduce the search time & memory time. So my focus is on this problem statements to enhance the performance of the Modified Apriori algorithm. After completing literature review I concluded that this problems can be solved by combining a clustering technique with the Modified Apriori algorithm. After finishing review of all clustering techniques in data mining I found the OPTICS (Ordering points to identify clustering structure) as best choice to solve this problem set.

V. OPTICS : DENSITY BASED CLUSTERING ALGORITHM

1) *Motivation :* Apriori is revision in plenty of research papers. Partition based clustering & Graph based clustering is also widely used for clustering of data. But density based clustering has less documentation in literature review than others. These clustering schemes are becoming more of interest among researchers. Upon reviewing every mining techniques I found the Optics as best solution for this enhancement task.

2) *Density based clustering :* Able to find clusters with arbitrary shapes. Dense region of objects in data set are referred as a cluster & are separated by region of low density. Cluster analysis is method of grouping a set of objects such that object having more similarity remains in same cluster & other than different cluster. Widely used for application areas like machine learning , pattern discovery , information retrieval & biomedical data. DBSCAN , OPTICS, DENCLUE etc are example of density based clustering.

3) *OPTICS :* Ordering points to identify the clustering structure. It gives cluster of data with ordering. Order points by shortest reachability distance to guarantee that clusters w.r.t. higher density are finished first. Based on this idea two values need to be stored for each objects : core distance and reachability distance[4].

```
OPTICS(DB, eps, MinPts)
   for each point p of DB
      p.reachability-distance = UNDEFINED
   for each unprocessed point p of DB
      N = getNeighbors(p, eps)
      mark p as processed
      output p to the ordered list
      if (core-distance(p, eps, Minpts) != UNDEFINED)
         Seeds = empty priority queue
         update(N, p, Seeds, eps, Minpts)
         for each next q in Seeds
            N' = getNeighbors(q, eps)
            mark q as processed
            output q to the ordered list
            if (core-distance(q, eps, Minpts)
                            != UNDEFINED)
               update(N', q, Seeds, eps, Minpts)
```

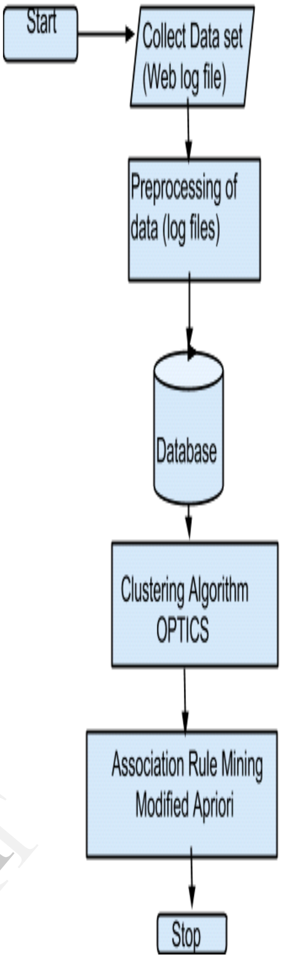
```
update(N, p, Seeds, eps, Minpts)
  coredist = core-distance(p, eps, MinPts)
  for each o in N
     if (o is not processed)
        new-reach-dist = max(coredist, dist(p,o))
        if (o.reachability-distance == UNDEFINED)
           // o is not in Seeds
           o.reachability-distance = new-reach-dist
           Seeds.insert(o, new-reach-dist)
        else              // o in Seeds, check for improvement
           if (new-reach-dist < o.reachability-distance)
              o.reachability-distance = new-reach-dist
              Seeds.move-up(o, new-reach-dist)
```

It can discover the clustering groups with irregular shape, uncertain amount and noises. Final clustering structures are insensitive to parameters. It can discover high density data included in low density groups. So it is preferable to choose this algorithm to solve the problem stated above. The following is pseudocode for algorithm.

OPTICS Outputs order of points, core-distance of points, reachability-distance of points

VI. PROPOSED APPROACH - MAPTICS

This research work proposed an novel approach for web usage mining process using cluster implementation. MAPTICS stands for Modified Apriori with OPTICS. This will overcome the drawbacks of existing algorithm & also enhances its performance. MAPTICS is fundamentally different concept than older ones. This combines the strength of OPTICS to that Modified Apriori explained above. Here, I am displaying flowchart of MAPTICS.


Start
Collect Data set (Web log file)
Preprocessing of data (log files)
Database
Clustering Algorithm OPTICS
Association Rule Mining Modified Apriori
Stop

Flowchart of MAPTICS

The first step of algorithm includes collection of web log file and performing pre-processing operation. Preprocessing stage includes task of extracting all valid requests ,user & session identification, data cleansing. After completion of preprocessing web log file is stored in database. The second step uses OPTICS clustering algorithm to find common behavior in data set & finds common access patterns. The third step covers association rule mining i.e. Modified Apriori to complete the process of pattern discovery.

When large data set is to be processed, directly processing it is complex but dividing its data into clusters will make the algorithm performance more better. So there need of cluster of data & finding frequent item sets

VII. CONCLUSION

Web Usage Mining techniques are great area of research since long. Aim is to get personalization , system improvement , adjustment of web site , business intelligence, improving e-commerce. In this paper , after reviewing 50+ research paper I have stated Modified Apriori algorithm with its problem set. I have chosen OPTICS algorithm as solution of it & proposed a new MAPTICS algorithm. This enhances the performance of existing one even for large data set.

FUTURE WORK

In future I will measure MAPTICS algorithm's performance with all the other similar kind of algorithms & results will be prepared. I will also expand this algorithm to solve real world problems.

REFERENCES

[1] Shaily G.Langhnoja , Mehul P. Barot , Darshak B. Mehta ,Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery, International Journal of Data Mining Techniques and Applications, ISSN: 2278-2419, Vol 02, Issue 01, June 2013

[2] http://commons.wikimedia.org/wiki/FileThe_general_relationship between_the_categories_of_Web_Mining_and_objectives_of_Data_Mining.png

[3] Bhaiyalal Birla*, Sachin Patel, International Journal of Advanced Research in Computer Science and Software Engineering , ISSN: 2277 128X , Volume 4, Issue 2, February 2014

[4] Rupinder Kaur, Simarjeet Kaur, A review: Techniques for clustering of web usage mining, International Journal of Science & Research (IJSR) , ISSN : 2310-7064, Volume 3 Issue 5, May 2014

[5] Sonali Manoj Raut, Dhananjay Dakhane , Comparative study of clustering & association method for large database in time domain, International journal of advance research in computer science & software engineering, ISSN : 2277 128x, Volume 2 , Issue 12 , December 2012