IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

Intelligent Frequent Pattern Analysis in Web Mining

Saurabh Bhattacharya
babu.saurabh@gmail.com

Dr.Sourabh Rungta
sourabh@rungta.ac.in

Naresh Kar
nareshkar@gmail.com

Abstract – **Web usage mining aims to discover interesting user access patterns from web logs. Web usage mining has become very critical for effective web site management, creating adaptive web sites, business and support services, personalization and so on. In this research, an efficient approach for frequent pattern mining using web logs for web usage mining is proposed and this approach is called as intelligent frequent pattern analysis. In this approach, the proposed technique is applied to mine association rules from web logs using normal Apriori algorithm, but with few adaptations for improving the interestingness of the rules produced and for applicability for web usage mining. Before mine the association rules, we are going to classify data with fuzzy clustering which is optimized through genetic algorithm. Association mining often produces large collections of association rules that are difficult to understand and put into action. In this research effective and intelligent pruning techniques have proposed that are characterized by the natural web link structures. Experiment results gives the interestingness measures can successfully be used to sort the discovered association rules after the pruning method was apply. Most of the rules that are rank highly according to the interestingness measures prove to be truly valuable to a web site administrator.**

Keywords – **Web usage mining, Apriori algorithm, Fuzzy Clustering, Association mining.**

I. INTRODUCTION

With the volatile growth of information available on the World Wide Web, several intelligent web services have been developed to help users for access relevant information from the Web. Though, the present web services provided are far from enough to satisfy the needs of diverse web users, particularly with the appearance of the Semantic Web. Consequently, further research needs to be carried out to identify new intelligent techniques and services for web users. To accomplish this, web mining is a promising direction to provide enhanced web services, mainly on the Semantic Web. When web sites are visited by users, web log files are generated on web servers and these files contain a huge amount of information. Data mining techniques like Clustering and Association Rule mining can be applied on this data to discover interesting information which when analysed by the web site maintenance engineer can reveal vital information required for web site improvement and there by attract more users to access the web site. Web usage mining has various application areas such as link prediction, web pre-fetching, web personalization and site reorganization.

Originally, association rule mining algorithms were applied for Market Basket Analysis which contained transaction data. The transaction data may include many records of which each record has a transaction id and a list of items purchased during that transaction. The main problem of web usage mining is to properly classify the data before mining association rules, so our main aim is to classify data efficiently before association rule mining.

According to analysis targets, web mining can be divided into three different types, which are:

A. Web Content Mining

Web content mining is the extraction and combination of valuable data, information and facts from Web page content. The heterogeneity and the absence of structure that permeates much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated innovation, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, WebCrawler, Alta Vista, Meta-Crawler, ALIWEB and others provide some comfort to consumers, but they do not usually provide structural information nor categorize, filter, or understanding documents.

Web content mining is differentiated from two different points of view: Information Retrieval View and Database View. A researcher summarized the research works done for unstructured data and semi-structured data from the information retrieval view. It demonstrates that most of the researches use bag of words, which is built on the statistics about single words in separation, to represent formless text and take single word found in the training corpus as features.

B. Web Structure Mining

Web structure mining is the process of using graph theory to analyse the node and connection structure of a website.

C. Web usage mining

Web usage mining is the process of extracting useful information from server logs e.g. Users' history [1]. It is the process of finding out what users are looking

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

for on the Internet. Certain users might be looking at only textual information, although some others might be interested in multimedia information. Web Usage Mining is the application of data mining techniques to invent interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data capture the identity or origin of Web users along with their browsing behaviour on a Web site.

The most important difficulty with web mining in common and web usage mining in specific is the temperament of the data they deal with. With the exception of the quantity of the data, the data is not absolutely structured. It is in a semi-structured arrangement hence it needs numerous preprocessing steps before the extraction of the essential information.

Web usage mining [1] aims to discover interesting and frequent user access patterns from web browsing data. The discovered knowledge can then be used for many practical web applications such as adaptive web sites, web references, and personalized web search and surfing.

Web Usage Data

Web usage data records access patterns of users from websites. However, it can also include data from user queries, registration information, user profiles, cookies, bookmarks, and any other data derived from the interactions of users while surfing on the Web. Web usage data are mainly divided into three categories, namely web server logs, proxy server logs and client browser logs.

II. REVIEW OF PREVIOUS WORK

Data in Web Usage Mining, can be obtained in server logs, browser logs, proxy logs, or collected from an organization's database. These data collections vary in terms of the location of the data source, the types of data accessible, the segment of population from which the data was obtained, and techniques of implementation [11].

This work going to make research in the field of Web usage mining (WUM), WUM is a division of Web Mining, which, sequentially, is a component of Data Mining. The process of mining significant and valuable information from vast database is called Data Mining [12]. WUM mines the usage features of the users of Web Applications. This obtained data can then be applied in a various ways such as, checking of fake elements etc.

WUM is considered as a component of the business intelligence in an organization [13]. It is applied for deciding business approaches via the competent use of web applications. It is very dynamic for the customer relationship management (CRM) since it can guarantee customer fulfilment till the interface between the customer and the organization is concerned [14].

Much research [1] has been carried out to mine web logs to discover interesting and frequent user access patterns. Sequential pattern mining techniques [6] are commonly used for discovering web access patterns from web logs. They are mainly based on two approaches: Apriori-based mining algorithms [4] such as AprioriAll [6] and GSP (Generalized Sequential Pattern mining algorithm) [7] and WAP-tree (Web Access Pattern Tree) based mining algorithms such as WAP-mine (WAP-tree mining) [8], FS-mine (Frequent Sequence tree mining algorithm) [10] and PLWAP (Pre-Order Linked WAP-Tre Mining) [9].

III. METHODOLOGY

Figure 1 shows the basic flow diagram for this research work.

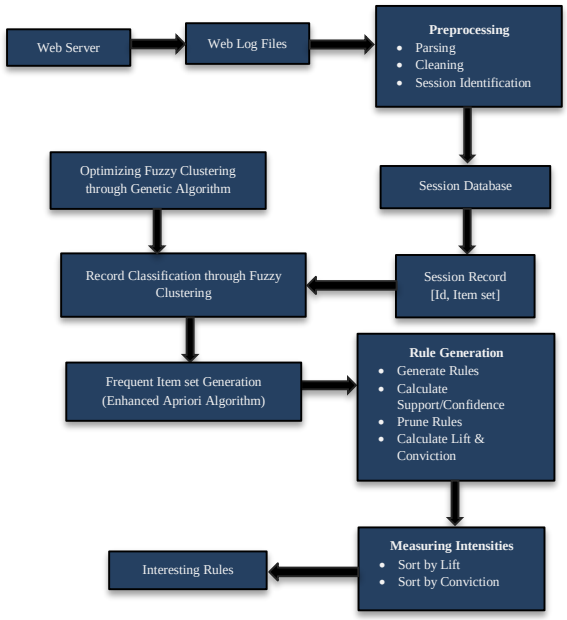

Figure 1: Flow diagram

Fuzzy C-Means Clustering

Objective function based fuzzy clustering algorithms such as the fuzzy c-means (FCM) algorithm has been used extensively for different tasks such as pattern recognition, data mining, and image processing and fuzzy modeling.

In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, maybe in the cluster

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

to a lesser degree than points in the centre of cluster. An overview and comparison of different fuzzy clustering algorithms are available.

With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)}$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The fuzzy c-means algorithm is very similar to the k-means algorithm:

1. Choose a number of clusters.
2. Assign randomly to each point coefficients for being in the clusters.
3. Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than, the given sensitivity threshold) :
   - Compute the centroid for each cluster, using the formula above.
   - For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes.

Algorithmic steps for Fuzzy c-means clustering

Let $X = \{x_1, x_2, x_3 \ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots, v_c\}$ be the set of centers.

1) Randomly select 'c' cluster center.

2) Calculate the fuzzy membership μ_{ij} using:

$$\mu_{ij} = 1/\sum_{k=1}^{c} (d_{ij}/d_{ik})^{(2/(m-1))}$$

3) Compute the fuzzy center v_j' using:

$$v_j = \frac{\left(\sum_{i=1}^{n} {\mu_{ij}}^m x_i\right)}{\sum_{i=1}^{n} {\mu_{ij}}^m} \quad \forall_j = 1,2, \ldots c$$

4) Repeat step 2) and 3) until the minimum 'J' value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

Where,

'k' is the iteration step.

'β' is the termination criterion between [0, 1].

'$U = (\mu_{ij})_{n*c}$' Is the fuzzy membership matrix.

'J' is the objective function.

Genetic Algorithm

A genetic algorithm is a probabilistic search technique that computationally simulates the process of biological evolution. It mimics evolution in nature by frequently altering a population of candidate solutions until an optimal solution is found.

The GA evolutionary cycle starts with a randomly selected initial population. The changes to the population happen through the processes of selection based on fitness, and alteration using mutation and crossover. The application of selection and alteration leads to a population with a higher proportion of improved solutions. The evolutionary cycle carry on until an acceptable solution is found in the current generation of population, or some regulator parameter such as the number of generations is exceeded.

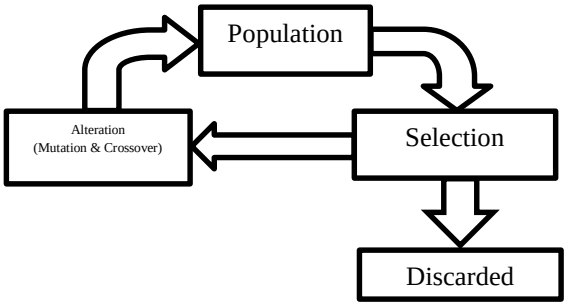

Figure 2: Genetic algorithm evolutionary cycle

The smallest unit of a genetic algorithm is called a gene, which denotes a unit of information in the problem domain. A series of genes, recognized as a chromosome, signifies one possible solution to the problem. Each gene in the chromosome signifies one component of the solution pattern.

The most common form of representing a solution as a chromosome is a string of binary digits. Each bit in this string is a gene. The procedure of converting the solution from its unique form into the bit string is known as coding. The specific coding system used is application dependent. The solution bit strings are cracked to enable their evaluation using a fitness measure.

A. Selection

In biological evolution, only the fittest survive and their gene pool contributes to the creation of the succeeding generation. Selection in GA is also based on a similar process. In a common form of selection, recognized as fitness proportional selection, every chromosome's likelihood of being selected as a decent one is proportional to its fitness value.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

B. Alteration to improve good solutions

The alteration step in the genetic algorithm refines the good solution from the current generation to produce the next generation of candidate solutions. It is takes place by acting out crossover and mutation.

C. Crossover

Crossover may be regarded as artificial mating in which chromosomes from two individuals are combined to create the chromosome for the next generation. This is carried out by splicing two chromosomes from two different solutions at a crossover point and swapping the spliced parts. The fact is that some genes with good characteristics from one chromosome may as a result combine with some good genes in the other chromosome to create a better solution represented by the new chromosome.

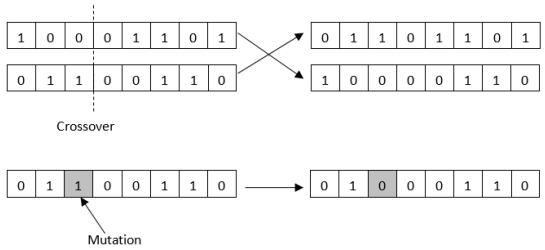

Figure 3: block representation of Crossover and Mutation

D. Mutation

Mutation is a random adjustment in the genetic composition. It is beneficial for announcing new characteristics in a population – something not achieved through crossover alone. Crossover only reorders prevailing characteristics to give new combinations. For instance, if the first bit in every chromosome of a generation happens to be a 1, any novel chromosome created through crossover will also have 1 as the first bit.

The mutation operator changes the current value of a gene to a different one. For bit string chromosome this modification amounts to flipping a 0 bit to a 1 or vice versa. Mutations can be counterproductive, and applied only randomly and infrequently.

The steps in the typical GA for finding a solution to a problem are listed below:

1. Generate an initial solution population of a certain size randomly.
2. Calculate each solution in the current generation and assign it a fitness value.
3. Select "good" solutions based on fitness value and discard the rest.
4. If satisfactory solution(s) found in the current generation or maximum number of generations is exceeded then stop.
5. Change the solution population using crossover and mutation to create a new generation of solutions.
6. Go to step 2.

Apriori Algorithm

Apriori is a classic algorithm which is proposed by Agrawal & Srikant [4] for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. The problem is that we are given a set of items and a large collection of transactions which are sets (baskets) of items. The task is to find relationships between the containments of various items within those baskets.

It is an iterative approach and there are two steps in each iteration. The first step generates a set of candidate item sets. Then, in the second step we count the occurrence of each candidate set in database and prune all disqualified candidates (i.e. all infrequent item sets). Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent , all its subset should be in last frequent item set.

The iterations begin with size 2 item sets and the size is incremented after each iteration. The algorithm is based on the closure property [4] of frequent item sets: if a set of items is frequent, then all its proper subsets are also frequent.

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ.

Initialize: k := 1, C_1 = all the 1- item sets;
read the database to count the support of C_1 to determine L_1.
L_1 := {frequent 1- item sets};
k:=2; //k represents the pass number//
while ($L_{k-1} \neq \emptyset$) do
begin
C_k := gen_candidate_itemsets with the given L_{k-1}
prune(C_k)
for all transactions $t \in T$ do
increment the count of all candidates in C_K that are contained in t;
L_k := All candidates in C_k with minimum support ;
k := k + 1;
end
Answer := $\cup_k L_k$

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

IV. SIMULATION AND RESULTS

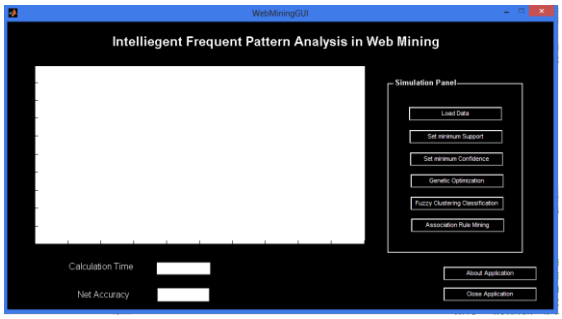

Figure 4: Main Graphical user Interface (GUI) for the simulation

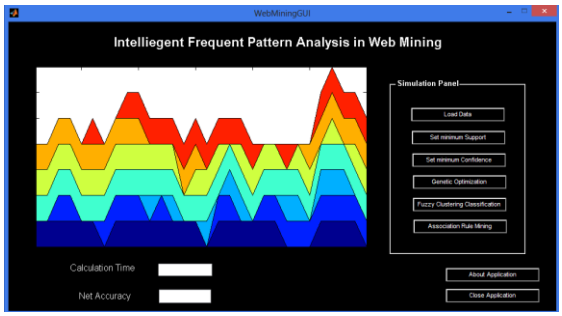

Figure 5: View of User Interface for data input

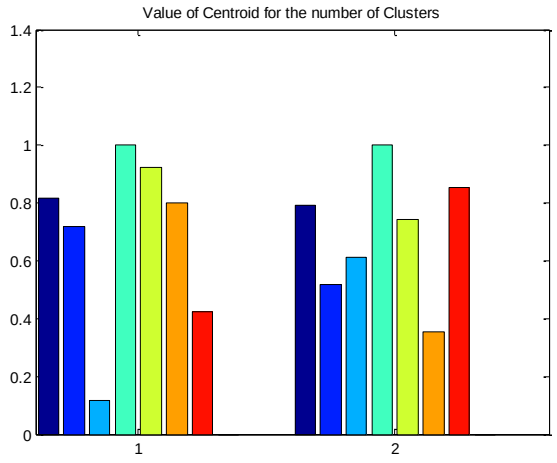

Figure 6: Matrix of final cluster center where each row provides the center coordinates

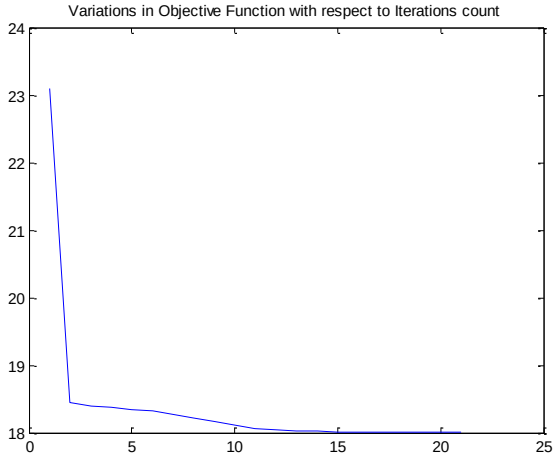

Figure 7: Variations in the values of the objective function during iterations

V. CONCLUSION

The candidate generation and test approach outperforms the pattern-growth approach on mining short patterns, while pattern-growth approach is better on mining long patterns. This research deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of this paper is to introduce the process of web log mining, and to demonstrate how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behaviour. We done experiment on Msnbc.com's weblog data, this data describes the page visits of users who visited msnbc.com on September 28, 1999. Visits are documented at the level of URL category and are recorded in time order. Results are recorded in terms of net accuracy and computational efforts and are satisfactory as compare with previous researches.

REFERENCE

[1] Srivastava J., Cooley R., Deshpande M., and Tan P. N., "Web Usage Mining: Discover and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations, Vol. 1, No. 2, pp. 12-23, 2000.

[2] Berners-Lee T., Hendler J., and Lassila O., "The Semantic Web", In Scientific American, May 2001.

[3] Cooley R., Mobasher B., and Srivastava J., "Data Preparation for Mining World Wide Web Browsing Patterns", In Journal of Knowledge and Information Systems, Vol. 1, Issue No. 1, 1999.

[4] Agrawal R. and Srikant R. "Fast Algorithms for Mining Association", In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp. 487-499, 1994.

[5] Han J., Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation", In Proceedings of the ACM SIGMOD International Conference on the Management of Data, pp. 1-12, 2000.

[6] Agrawal R., and Srikant R., "Mining Sequential Patterns", In Proceedings of the 11th International

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary research
Website: www.ijdacr.com (Volume 2, Issue 3, October 2013)

Conference on Data Engineering, Taipei, Taiwan, pp. 3-14, 1995.

[7] Srikant R., and Agrawal R., "Mining Sequential Patterns: Generalizations and Performance Improvements", In Proceedings of the 5th International Conference on Extending Database Technology (EDBT), Avignon, France, pp. 3-17, 1996.

[8] Pei J., Han J., Mortazavi-asl B., and Zhu H., "Mining Access Patterns Efficiently from Web Logs", In Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Kyoto, Japan, pp. 396-407, 2000.

[9] Lu Y., and Ezeife C.I, "Position Coded Pre-order Linked WAP-Tree for Web Log Sequential Pattern Mining", In Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Seoul, Korea, 2003.

[10] Maged E., Elke A. R., and Carolina R., "FS-Miner: An Efficient and Incremental System to Mine Contiguous Frequent Sequences", Computer Science Technical Report Series, Worcester Polytechnic Institute, 2003.

[11] Dr. G. K. Gupta, "Introduction to Data Mining with Case Studies", PHI Publication, 2005.

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, No. 2, Pp. 12-23, 2000.

[13] Adel T. Rahmani and B. Hoda Helmi, "EIN-WUM an AIS based Algorithm for Web Usage Mining", Proceedings of GECCO'08, Atlanta, Georgia, USA, ACM978-1-60558-1309/08/07, pp. 291-292, 2008.

[14] Shailey Minocha, Nicola Millard, Lisa Dawson, "Integrating Customer Relationship Management Strategies in (B2C) E-Commerce Environments", IFIP Conference on Human Computer Interaction-INTERACT, 2003.