ISSN: 2321-7782 (Online)
Volume 4, Issue 4, April 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Fuzzy c-means clustering algorithm implementation in User Behaviour Analysis

N. Pushpalatha[1]
Assoc. Professor in CSE
Marri Laxman Reddy Institute of Technology & Management
Hyderabad – India

K. Ram Chandra Reddy[2]
Dept. of CSE
Marri Laxman Reddy Institute of Technology & Management
Hyderabad – India

Abstract: ***In recent years the development of web had generated a massive data which is related to respective activity. A new method appears to analyses this data and they were grouped under the generic term of "web mining". Web mining is a technique which is used for research of data in web. It is also called weblog mining in this data mining techniques are applied for web access.***

Web mining is the application of artificial intelligence, data mining, chart technology and etc on the data which is in web and traces the user visiting and behaviour of their interest on the data. In many of the web services like e-commerce, analytic web, e-learning, information retrieval etc, in computing and information science web mining has become one of the important area for data retrieval. Mining methods are used in log data to extract the behaviour of the users in which the user searched the data in many applications like websites, personalized services etc.

Keywords: Clustering, Web Usage Mining, WWW, Web logs, pre-processing, Data Warehouse.

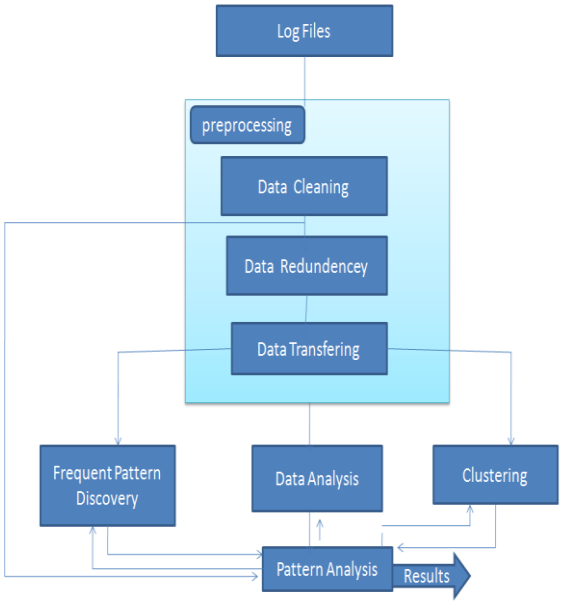
I. INTRODUCTION

The development of the Web that occurred in the recent years generated a boom of data related to its activities. To analyze (or rather excavate) these new types of data new methods appeared and were grouped under the generic term of "Web Mining". Web usage mining is an active, technique used in this field of research. It is also called web log mining in which data mining techniques are applied to web access log. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces users visiting behaviours and extracts their interests using patterns.

Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval etc., web mining has become one of the important areas in computer and information science. Web Usage Mining uses mining methods in log data to extract the behaviour of users which is used in various applications like personalized services, adaptive web sites, prefetching, creating attractive web sites etc.

II. PROPOSED ARCHITECTURE

System model describes about the functioning of the **WMS** system. History records taken as input then pre-processing is applied on the data. Then analysis is done for clustering to get the result.



III. LOG FILES

In computing, a logfile is a file that records either events that occur in an operating system or other software runs, or messages between different users of a communication software. Logging is the act of keeping a log. In the simplest case, messages are written to a single logfile.

A transaction log is a file of the communications between a system and the users of that system, or a data collection method that automatically captures the type, content, or time of transactions made by a person from a terminal with that system. For Web searching, a transaction log is an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine.

Many operating systems, software frameworks, and programs include a logging system. A widely used logging standard is S, defined in Internet Engineering Task Force The Syslog standard enables a dedicated, standardized subsystem to generate, filter, record, and analyze log messages. This relieves software developers of having to design and code their own ad hoc logging systems.

IV. INTERNET RELAY CHAT

In the case of IRC software, message logs often include system/server messages and entries related to channel and user changes making them more like a combined message/event log of the channel in question, but such a log isn't comparable to a true IRC server event log, because it only records user-visible events for the time frame the user spent being connected to a certain channel.

V. INSTANT MESSAGING

The use of data stored in transaction logs of Web search engines, Intranets, and Web sites can provide valuable insight into understanding the information-searching process of online searchers. This understanding can enlighten information system design, interface development, and devising the information architecture for content collections.

VI. PRE PROCESSING

Major Tasks in Data Pre-processing

- Data cleaning

- Data integration
- Data transformation
- Data reduction

VII. Data Cleaning

Data cleaning Real world data tend to be in complete, noisy and inconsistent .data cleaning routines attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Ways for handling missing values: a. Ignore the tuple: this is usually done when class label is missing. This method is not very effective, unless tuple contains several attributes with missing values.

It is especially poor when the percentage of missing values per attributes varies considerably. b. Fill in the missing value manually: this approach is time consuming and may not be feasible given a large data set with missing values. c. Use a global constant to fill in the missing value: replace all missing attribute values by the same constant, such as label like “unknown”. if missing value are replaced by ,say unknown then the mining program may mistakenly think that they form an interesting concept.

VIII. Data Integration and Transformation

Data integration It is likely that your data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are a number of issues to consider during data integration. Schema integration can be tricky. How can like real-world entities from multiple data sources be “matched up"? This is referred to as the entity identification problem. For example, how can the data analyst or the computer is sure that customer _id in one database, and cust_number in another refer to the same entity? Databases and data warehouses typically have metadata - that is, data about the data. Such metadata can be used to help avoid errors in schema integration.

Redundancy is another important issue. An attribute may be redundant if it can be “derived" from another attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis. Data transformation In data transformation, the data are transformed or consolidated into forms appropriate or mining.

Data transformation can involve the following a. Smoothing: which works to remove the noise from data. Such techniques include binning, clustering, and regression b. Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities c. Generalization of the data: where low level or “primitive" (raw) data are replaced by higher level concepts through the use of concept hierarchies.

For example, categorical attributes, like street, can be generalized to higher level concepts, like city or county. Similarly, values for numeric attributes, like age, may be mapped to higher level concepts, like young, middle aged, and senior d. Normalization: where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0. e. Attribute construction: where new attributes are constructed and added from the given set of attributes to help the mining process.

IX. Data Reduction

Imagine that you have selected data from the all electronics data warehouse for analysis. The data set will likely be huge! Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or

infeasible. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same analytical results. In this section, we first present an overview of data reduction strategies, followed by a closer look at individual techniques.

X. Frequent Pattern

In the first pass, the algorithm counts occurrence of items in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root. Recursive processing of this compressed version of main dataset grows large item sets directly, instead of generating candidate items and testing them against the entire database.

Growth starts from the bottom of the header table (having longest branches), by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree. Once the recursive process has completed, all large item sets with minimum coverage have been found, and association rule creation begins.

XI. Data Analysis

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis, and confirmatory data analysis.

EDA focuses on discovering new features in the data and on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis. Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination. The term data analysis is sometimes used as a synonym for data modeling.

XII. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics and data compression. Cluster analysis itself is not one specific algorithm, but the general task to be solved.

It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and

intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure.

XIII. EXPECTED RESULT

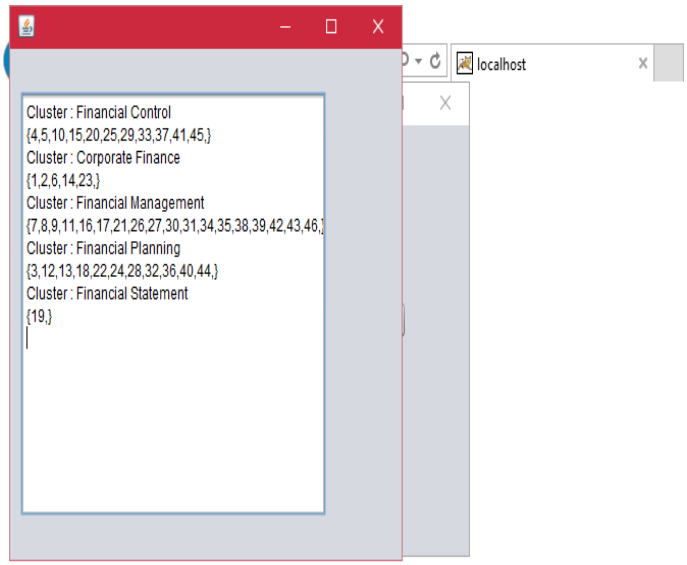


Figure: Expected output 1

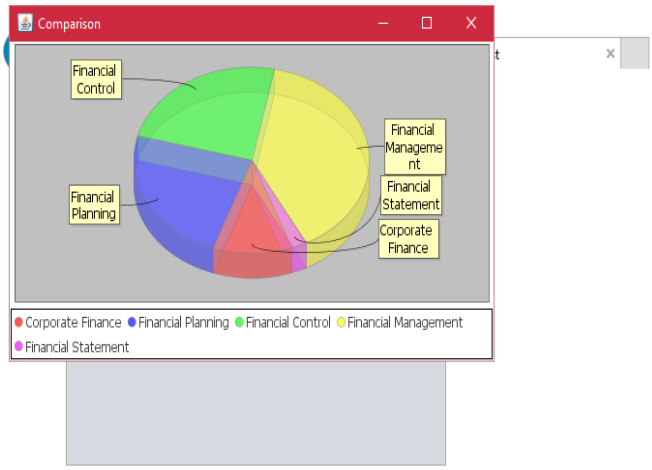


Figure:Expected output 2

XIV. CONCLUSION

In this project proposed methodologies used for classifying the user using Web Usage data. This model analysis the users behaviours and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. The results of evaluations shows that using more efficient algorithms for finding similar users lead to recommender system that provides more interesting recommendations for website users. Proposed work can be extended by considering the effect of users" feedback for increasing the quality of recommendation. This can be done, eventually, by introducing new parameters for the characterization of the Web Usage data.

References

1. M. Spiliopoulou, L. C. Faulstich, and K. Winkler. A data miner analyzing the navigational behaviour of web users. In Proc. of the Workshop on Machine Learning in User Modeling of the ACAI'99 Int. Conf., Creta, Greece, July 1999.
2. M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In AAAI/IAAI, pages 727{732, 1998.
3. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model-based clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 280{284, Boston, Massachusetts, 2000.
4. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery, 6(1):61{82, January 2002.
5. F. Bonchi, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, and S. Ruggieri. Web log data warehousing and mining for intelligent web caching. Data Knowledge Engineering, 39(2):165{189, 2001.
6. Osmar R. Zaiane, Man Xin, and Jiawei Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Advances in Digital Libraries, pages 19-29, 1998.
7. F. SÄauberlich and K.-P. Huber. A framework for web usage mining on anonymous logfile data. In Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the Gesellschaft fÄur Klassifikation e.V., March 2001, pages.

