IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Web Content Mining – Survey

M.Santhanakumar[1] and Dr.P.Eswaran[2]

[1]**Department of CSE, PSN College of Engineering and Technology, Tirunelveli, Tamilnadu - 627152, India**

[2]**Department of CSE, PSN College of Engineering and Technology, Tirunelveli, Tamilnadu - 627152, India**

Abstract

Due to rapid increase in the information and human needs, which leads to the new inventions, we need huge amount of repositories in computers for storing and retrieving the information whenever needed. That information may shared with some other people through the World Wide Web. Web browsers are the tools, used to bring the information in any fields to the users. Web mining, the term can be defined as repossessing the information and also extracting the human knowledge through the web. Retrieving the information from the web has become one of the most challenging task.Web mining, basically divided into three major categories 1) Web Content Mining 2) Web Usage Mining 3) Web Structure Mining. Web Content mining is defined as Extraction and Integration of useful data, information and Knowledge from web page contents. This paper contains the concepts and details of web content mining and various clustering algorithms to be used for Web Content Mining.

Keywords: *Web Mining, Web Content Mining, Web Usage Mining and Web Structure Mining, Clustering.*

1. Introduction

Data mining [Chen et al. 1996] is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery in Database (KDD). The KDD processes are shown in Figure1. Usually there are three processes.

Pre-processing - This is executed before data mining techniques are applied to the right data. The pre-processing includes data cleaning, integration, selection and transformation.

Data mining process – In this process different algorithms are applied to produce hidden knowledge.

Post-processing - This evaluates the mining result according to user's requirements and domain knowledge.

The data sources may come from different databases, which may have some inconsistence and duplicate data. So first we need to clean and integrate the databases by removing noises from the data source. The second task is to select related data from the integrated resources and transform them into a format that is ready to be mined. After selection of relevant data, the database in which our data mining techniques are to applied will be much smaller, consequently the whole process will be more efficient.

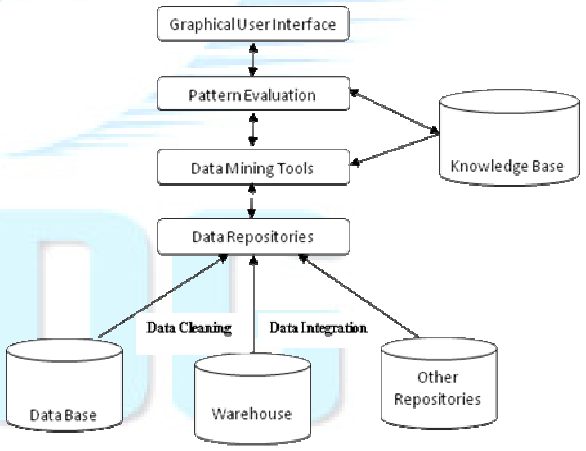

Figure 1: Knowledge Discovery Process

2. Types of Data

2.1 Relational database

Relational database is highly structured data repository, where the data are described by a set of

attributes and stored in tables. With the well developed database query languages, data mining on relational database is not difficult. Data mining on relational database mainly focuses on discovering patterns and trends.

2.2 Transactional database

Transactional database refers to the collection of transaction records. In most cases they are sales records. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in the transaction records.

2.3 Spatial database

Spatial databases usually contain not only traditional data but also the location or geographic information about the corresponding data. Spatial association rules describe the relationship between one set of features and another set of features in a spatial database.

2.4 Temporal and time-series database

Differ from traditional transaction data. For each temporal data item the corresponding time related attribute is associated. Temporal association rules can be more useful and informative than basic association rules.

2.5 World-Wide Web

As information on the web increases in a phenomena speed and web becomes ubiquitous, most researchers turn to the field of mining web data (Web Mining). Web mining is usually divided into three main categories.

Web usage mining, concentrates on mining the access patterns of users.

Web structure mining, focuses in analysis of *structures* and links in web documents.

Web content mining, includes text mining, multimedia mining and graphic mining.

3.Types of Mining

1. Predictive data mining, which produces the model of the system described by the given data set.

2. Descriptive data mining, which produces new, nontrivial information based on the available data set.

4. The Primary Data Mining Tasks

4.1 Classification - discovery of a predictive learning function that classifies a data item into one of several predefined classes.

4.2 Regression - discovery of a predictive learning function, which maps a data item to a real-value prediction variable.

4.3 Clustering - a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.

4.4 Summarization - an additional descriptive task that involves methods for finding a compact description for a set (or subset) of data.

4.5 Dependency Modeling - finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set.

4.6 Change and Deviation Detection - discovering the most significant changes in the data set.

5.Clustering Algorithms

5.1 Hierarchical Clustering (Connectivity Based Clustering):

[1]Hierarchical techniques produce a nested sequence of partitions, with a single and all inclusive cluster at the top and singleton clusters of individual objects at the bottom. Clusters at an intermediate level encompass all the clusters below them in the hierarchy. The result of a hierarchical clustering algorithm can be viewed as a tree, called a dendrogram. That is shown in figure 2.

Based on the direction of building the dendrogram of a hierarchical clustering two major categories can be classified they are,

5.1.1 Agglomerative – In this each observation starts from its own cluster and pairs of clusters that are merged as one while moving up the hierarchy (Bottom up). It is most commonly used technique in Hierarchical Clustering.

5.1.2 Divisive – In this all observations starts in a single cluster and splits are done recursively, while moving down the hierarchy (Top down).

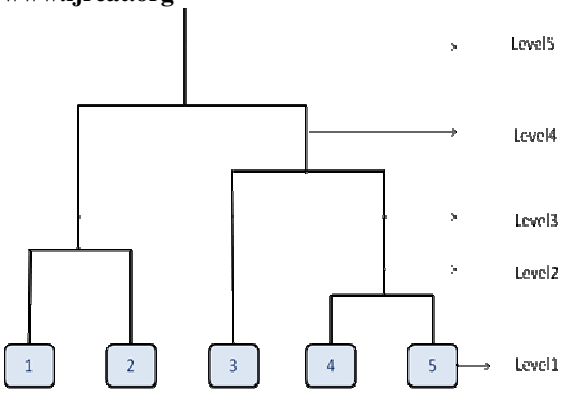

Figure 2: An Example Dendrogram of Hierarchical Clustering

Level 1 = {1},{2},{3},{4}.{5}
Level 2 = {1},{2},{3},{4,5}
Level 3 = {1,2},{3},{4,5}
Level 4 = {1,2},{3,4,5}
Level 5 = {1,2,3,4,5}

5.2 Bidirectional Hierarchical Clustering

[2] Proposed a new algorithm, bidirectional Hierarchical clustering system consisting of five major steps.
a. Representing web pages by vector space model
b. Generate the matrix of k-nearest neighbours of web pages
c. Bottom up cluster merging phase
d. Top-down refinement phase
e. Extracting the concepts of clusters

5.3 Nearest Neighbour Join Algorithm

This is the method used to perform several types of agglomerative hierarchical clustering algorithm. Here using an amount of memory that is linear in the number of points to be clustered in a group and an amount of time linear in the number of distinct distances between pairs of points. The main aim of this algorithm is to find the pairs of each clusters that is to be merged by the path in the nearest neighbour graph.

5.4 *k*-means Clustering: (Centroid based clustering)

In Centroid based clustering method central vectors are used to represent the clusters. In k-means clustering find k no of cluster centres and assign the objects (information) to the nearest cluster centre. This common approach is to search only for approximate solutions. It commonly runs multiple times with different random initialization to find an local optimum. Variations of k-means often include such optimizations for choosing the best of multiple runs.
The disadvantage here is the k-means algorithm needs the number of clusters (k) in advance and also the algorithm prefers the approximately similar size of clusters. It is also referred as Lloyd's algorithm.

5.5 *k*-medians Clustering

In this clustering algorithm calculate the median, instead of calculating mean value for each cluster. This algorithm relates directly to *k* median problem. Which is used to minimize the distance from each cluster to the nearest centers.

5.6 Divisive Hierarchical Clustering

[1] Illustrates the divisive hierarchical clustering that works from top to bottom of the cluster. Clustering starts with whole data set of one cluster and split the cluster step by step until an individual object is remained. There are two basic things which cluster split next and how to perform the splitting. While splitting the cluster it is to be decided that which object to take part in which sub-cluster.

5.7 Bisecting K-means

[3] In Bisecting K-means, it is started with a lone clustering of all documents. There are some steps to perform the above clustering they are,
1. Select a cluster
2. Using basic K-means clustering split the cluster into two sub clusters.
3. Repeat the step 2 for the fixed number of times till it generates the highest similarity of the clusters.
4. Repeat the above three steps until required clusters is reached.

6. Conclusion

In this paper, we classify and discuss about various types of databases and primary data mining tasks. Web Content mining is useful for establishing better relationship with customer by the way of providing information what they need. Here we also describe about different clustering algorithms used in Web Content Mining for effective personalization of information.

References

[1]Web Mining: Clustering Web Documents A Preliminary Review, Khaled M. Hammouda Department of Systems Design Engineering University of Waterloo Waterloo, Ontario, Canada N2L 3G1,February 26, 200.
[2]Bidirectional Hierarchical Clustering for Web Mining, Zhongmei Yao, Ben Choi, college of Engineering and science, Louisiana Tech University, Ruston, LA71272, USA

[3] A Comparison of Document Clustering Techniques, Michael Steinbach, George Karypis, Vipin Kumar, Department of Computer Science / Army HPC Research Center, University of Minnesota, 4-192 EE/CSci Building, 200 Union Street SE, Minneapolis, Minnesota 55455.

[4] Incremental Web Usage Mining Based on Active Ant Colony Clustering, SHEN Jie, LIN Ying, CHEN Zhimin, Wuhan University Journal of Natural Sciences, Vo1.11 No.5 2006 1081-1085, Article ID: 1007 1202(2006)05 1081-05.

[5] Web Content mining techniques: Survey, Faustina Johnson, Santosh Kumar Gupta, Department of Computer Science and Engineering, Krishna Institute of Engineering and Technology, Ghaziabad, International Journal of Computer Applications (0975—888) Volume 47-No.11, June 2012.

[6] Partition Based Web Content Mining, Niki R.Kapadia, Kanu Patel, Mehul C.Parikh, GEC, Modasa, International Journal of Engineering Research and Technology, ISSN:2278-0181, Vol 1 Issue 3, May 2012.