Event Sequence Analysis using Self Organizing Map

Dharmendra Kaushik, Babita Kubde
1M.Tech Scholar, 2Reader, PHD Scholar
#Dept. Of Computer Science & Engineering RCET, Bhilai
Chhattisgarh, Bhilai - India
dharam.rcm@gmail.com
babita_g7@rediffmail.com

ABSTRACT

In today’s world we have abundance of data and scarcity of Knowledge data mining field emerged as the fit of the tool to the problem. With the advent of internet technology and the exponential growth in the technology behind the world wide web, the concept of web mining and found a place for itself and emerged as a separate field of research. Web mining involves a wide range of applications that aim at discovering and extracting hidden information in data stored on the Web. Web log analysis is an innovative and unique field constantly formed and changed by the convergence of various emerging Web technologies. Due to its interdisciplinary character, the diversity of issues it addresses, and the variety and number of Web applications, it is the subject of many distinctive and diverse research methodologies. This chapter examines research methodologies used by contributing authors in preparing the individual chapters for this handbook, summarizes research results, and proposes new directions for future research in this area.

KEYWORDS - Data Mining, Web Mining, Preprocessing web log, weblog mining, web usage mining

INTRODUCTION

The expansion of the World Wide Web has resulted in a large amount of data that is now in general freely available for the user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the web is now the focusing area of an increasing number of researchers. Several data mining methods have been used to discover the hidden information in the Web. Web mining has been developed into an autonomous research area.

Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining.

Usually, raw data are stored by companies only for operational purpose and are not used to extract useful information for decision-making. These raw data are normally represented as a set of single spatial-temporal points, which is intuitively known as weblog. Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

The advent of new generation of techniques called 'Data Mining' has brought about opportunities to explore knowledge from patterns in huge amount of data. Data mining on web data is termed as ‘Web Mining’. In this work, it has been tried to explore web mining techniques that have contributed towards achieving atomization in web page delivery to a web surfer. In my opinion, Personalization can be achieved in two steps –
1. By understanding the characteristics of surfer about his task and motive.
2. By organizing the web structure and content to deliver the appropriate or most relevant pages to the user.

In this paper, we are focusing on the weblog mining that means whatever the events occurred in the web by the user has been monitored and calculate the above by using clustering algorithm. Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of click stream data in usage logs in order to produce the right data for mining.

CLUSTERING

Cluster analysis is a class of statistical techniques that can be applied to data that exhibit “natural” groupings. Cluster analysis sorts through the raw data and groups them into clusters. A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters. Cluster analysis is the obverse of factor analysis. Whereas factor analysis reduces the number of variables by grouping them into a smaller set of factors, cluster analysis reduces the number of observations or cases by grouping them into a smaller set of clusters.

Types of Clustering

Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitioned algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold.

WEBLOG CLUSTERING

Clustering creates an illusion — it permits the deployment of application components and services to several machines while presenting only a single face to the client. There are good reasons to support this illusion. When a client requests a service, it should make no difference if the service runs on a single server or across a number of servers. The clustering abstraction provides you with a clear route to improving the performance and scalability of your applications, albeit with increased administration of hardware and network resources. Weblog clustering offers three important benefits:

Scalability
A solution that allows you to create additional capacity by introducing more servers to the cluster, thereby reducing the load on existing servers.

Load balancing
The ability to distribute requests across all members of the cluster, according to the workload on each server.

High availability
A mix of features that ensure applications and services are available even if a server or machine fails. Clients can continue to work with little or no disruption in a highly available environment. WebLogic achieves high availability using a combination of features: replication, failover, and migratable services.

PROPOSED METHODOLOGY

After each user has been identified (through cookies, logins, or IP/agent analysis), the click-stream for each user must be divided into sessions. As we cannot know when the user has left the Web site, a timeout is often used as the default method of breaking a user’s click-stream into sessions. Methods and algorithms used here have been developed from several fields such as statistics, machine learning, and databases. This phase of Web usage mining has three main operations of interest: association (i.e. which page tend to be accessed together), clustering (i.e. finding groups of users, transactions, pages, etc.), and sequential analysis (the order in which web pages tend to be accessed).
RESEARCH METHODOLOGY

What are the research methodologies frequently applied in Web-based research? Some researchers focus on collection and preparation of information for data analysis (Jansen, 2006), while others concentrate on elicitation; reduction and visualization for user-profiling (Romano et al., 2003). Researchers also benefit from a new, aggressively growing source of personal communication – blogs (Jing, 2006; Rossler, 2002). In a different direction, there are a number of studies that focus on analysis of research methodologies. Powel (1999) uses a comprehensive classification developed by Kim (1996) to review, define and discuss quantitatively and qualitatively-driven methodologies. Another publication (Palvia et al., 2007) provided a slightly different but equally comprehensive classification of research methodologies. Using these sources, we identified the following methodologies:

- **Conceptual Framework / Inquiry**: Concepts are introduced and defined, and subsequently used to construct conceptual frameworks that provide study directions.
- **Phenomenology / Ethnomethodology**: An interpretive methodology that examines users behavior. Ethnomethodology, an extension of phenomenology, examines individual and group interactions within a social structure.
- **Content Analysis**: A methodical and replicable methodology used to determine, quantify, and analyze the presence of research objects within a large data set.
- **Ethnography**: A qualitative study in which the researcher observes members of a chosen group in a natural environment over a long period of time.
- **Historical Method**: Collects and examines facts about events, people and the environment of the past.
- **Discourse Analysis**: A scientific argument evaluation method.
- **Case Study**: A comprehensive study of a single subject, influenced by a proper selection of unit of analysis.

**Clustering Parameters** The clusters were formally defined above, but it is difficult to formally evaluate the quality of a cluster. Many clustering algorithms are evaluated through a visual representation. In a similar way, we use a visual representation of the weblog and present the clustering result in a visual representation as well. Obviously clusters are important in order to show the quality of the clustering method. However, these clusters should not be obviously delimited in order to evaluate the influence of the clustering parameters in their quality.

As the requirement of our algorithm we need some basic parameters which are:

![Clustering Parameters](image)

**Inputs Data Set**

Clustering Experiments

The Self-Organizing Map (SOM) has proven to be one of the most powerful algorithms in data visualization and exploration. Application areas include various fields of science and technology, e.g., complex industrial processes, telecommunications systems, document and image databases, and even financial applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data.

Self-organization map feature:

- Dimensionality reduction of unsupervised learning.
- Can applied in deal huge amounts of sample.
- The original data set is represented using a smaller set of prototype vectors not to find an optimal clustering but to get good.

After selecting the limited subset of weblog and specifying required parameters we have performed SOM clustering in our data, to find a cluster, SOM starts with an arbitrary point p and retrieve all points density reachable from p wrt. Eps and MinPts. If p is a core point, this procedure yields a cluster wrt. Eps and MinPts. If p is a border point, no points are density-reachable from p and SOM visits the next point of the database.

Since we use global values for Eps and MinPts, SOM may merge two clusters into one cluster, if two clusters of different density are “close” to each other. Let the distance between two sets of points $S_1$ and $S_2$ be defined as $\text{dist}(S_1, S_2) = \min \{\text{dist}(p, q) | p \in S_1, q \in S_2\}$. Then, two sets of points having at least the density of the thinnest cluster will be separated from each other only if the distance between the two sets is larger than Eps. Consequently, a recursive call of SOM may be necessary for the detected clusters with a higher value for MinPts. This is, however, no disadvantage because the recursive application of SOM yields an elegant and very efficient basic algorithm. Furthermore, the recursive clustering of the points of a cluster is only necessary under conditions that can be easily detected.
As shown in fig we clearly identify that data set which are presented in one color set are belongs to one cluster and also represent the interest of a person in one place at one time and the fig given below is specify the number of dataset in a cluster.

EXPERIMENTAL RESULTS

This chapter presents the experiments performed to validate the proposed method. In these experiments we used weblog data collected by Microsoft Corporation.

A common need in analyzing large quantities of data is to divide the data set into logically distinct groups, such that the objects in each group share some property that does not hold (or holds much less) for other objects. As such, clustering searches a global model of data, usually with the main focus on associating each object with a group (i.e. a cluster), even though in some cases we are interested (also) in understanding where clusters are located in the data space. In this section, we focus on the context of moving objects and, thus, on the weblogs that describe their movement. In this setting, clustering consists essentially in trying to outline groups of individuals that show similar behaviors.

Moving in the same way at the same time is sometimes too restrictive to discover useful information, and thus the temporal constraint might be removed. In these cases, we could look for groups of objects that follow the same route but at any moment in time, thus formulating requests of the type

Cluster with three sub-clusters (right eye) spherical cluster (left eye) elliptical cluster (nose) non-spherical cluster (U-shaped: mouth) large and sparse cluster (body) noise (Such as black x).

Methods and Parameters:

Cluster step 1:
- Training Parameters of the SOM's Map size: 19x17
- Initial Neighborhood Widths: Rough Phases 1(0): 10
- Fine-Tuning Phases 2(0): 2
- learning rates:(The learning rate decreased linearly to zero during the training)
  - Rough Phases : 0.5 Fine-Tuning Phases 0.05

Cluster step 2:
- Method: K-Means Using 100 Runs

We conclude by mentioning the existence of other time series-based approaches that define distances between features extracted from series, rather than comparing the series themselves. For instance, we could extract all pairs of consecutive values in each series and then simply count the number of pairs shared by the two series compared.

CONCLUSION

This paper has presented the details of preprocessing tasks that are necessary for performing Click stream Data Mining in the application of data mining and knowledge discovery techniques for web log data collected by web servers. The contribution of the paper is to introduce the process of preparing web log data for different processes of click stream data.
mining, and then use web log for mining. The experimental results presented in section 4, illustrate the importance of the data preprocessing step and the effectiveness of our methodology, by reducing not only the size of the log file but also increasing the quality of the data available through the new data structures that we obtained. The process itself does not fully guarantee that we identify correctly all the transactions (i.e. user sessions & visits). This can be due to the poor quality of the initial log file Therefore, we need a solid procedure that guarantees the quality and the accuracy of the data obtained at the end of data preprocessing. In conclusion, our methodology is more complete because:

- It offers the possibility to analyze several click stream data.
- It employs the effective cleaning of unnecessary data.
- It gives the formatted log file in a text file format.

**FUTURE SCOPE**

Here we perform a simple laboratory based experiment which may be defer for real log, it may be modified some more realistic data. We are experimenting for more click stream log and realistic and sophisticated data. There may be some changes occurs on web log data. Some new phases may be introduced for time series data.

**REFERENCES**

7. Bettina Berendt, Web usage mining, site semantics, and the support of navigation, in Proceedings of the Workshop “WEBKDD’2000 - Web Mining for E-Commerce - Challenges and Opportunities”, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000, Boston, MA

