ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Study of navigation patterns based on Clustering

Pawan B Mulay [*1], Dr Harish Mittal [2]

[*1] Computer Science Department, Mtech (Pursuing) SatPriya Institute of Engineering & Technology, Rohtak, India

[2] Director, SatPriya Institute of Engineering & Technology, Rohtak, India

mulaypb@gmail.com

Abstract

The WWW continues to grow at an amazing rate as an information gateway and as a medium for business. Web mining extracts interesting and useful knowledge and information from artifacts or activity related to the WWW. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help understand the user behavior and the web structure. User profiles could be built by combining user's navigation paths with other data features, such as page viewing time, hyperlink structure, and page content. What makes the discovered knowledge interesting had been addressed by several works. In this paper we study the web usage mining of the users navigation patterns with help of K-means clustering algorithm.

Keywords: Web mining, Web Usage mining, Clustering, K-means clustering.

Introduction

Web mining is the application of data mining techniques to extract knowledge from Web data, in which at least one of structure or usage (Web Log) data is used in the mining process. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories.

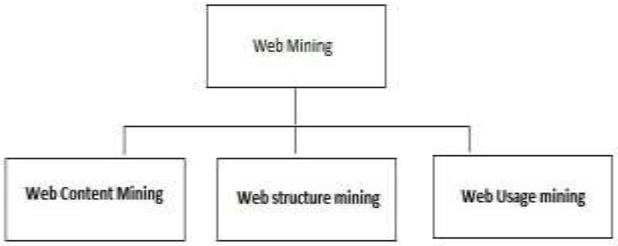

Fig 1: Classification of Web Mining[1]

1. **Web Content Mining**: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.

2. Web Structure Mining: The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks**:** A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan et al. provide an up-to-date survey.

Document Structure**:** In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

3. Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

be classified further depending on the kind of usage data considered:

***Web Server Data*:** The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

Application Server Data: Commercial application servers such as Web logic, Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

Web Usage Mining process is divided into three phases Pre-Processing, Pattern Discovery & Pattern Analysis as shown in figure below.

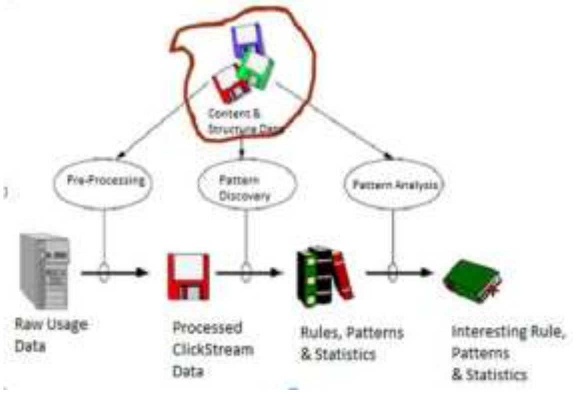
Content & Structure Data
Pre-Processing
Pattern Discovery
Pattern Analysis
Raw Usage Data
Processed ClickStream Data
Rules, Patterns & Statistics
Interesting Rule, Patterns & Statistics

Fig 2.Phases of Web Usage mining.[3]

a. Pre-processing Phase

The purpose of Data Pre-processing is to change a web data mining into reliable data. The normal procedure of data pre-processing includes 5 steps: data cleaning, user identification, user session identification, path completion and user transaction identification.

b. Pattern Discovery

It is used to find patterns using technique like.
Pathanalysis,Classification,Discovery,Clustering

c. Pattern Analysis

This phase uses techniques like.

- OLAP/ Visualization Tool: For Multidimensional analysis & Decision Making.
- Knowledge Query Management
- Intelligent Agents.

K-means clustering algorithm

K-means clustering method is a common division-based clustering method, also known as K-means method, is a widely used algorithm. A form of clustering will make an objective criteria for the classification (often referred to as the similarity function, such as: distance, similarity coefficient) optimization. The K-Means algorithm is one of a group of algorithms called partitioning clustering algorithm. The most commonly use partitioned clustering strategy is based on square error criterion. The general objective is to obtain the Partition that, for a fixed number of clusters, minimizes the total square errors. Suppose that the given set of N samples in an n-dimensional space has somehow been partitioned into K-clusters {C1, C2, C3… C_K}. Each C_K has n_K samples and each sample is in exactly one cluster, so that $\Sigma\ n_K = N$, where k=1…K. The mean vector M_k of cluster C_K is defined as the centroid of the cluster.

$$M_K = (1/n_k) \sum_{i=1}^{n_k} x_{ik}$$

Where x_{ik} is the i^{th} sample belonging to cluster C_K. The square-error for cluster C_K is the sum of the squared Euclidean distances between each sample in C_K and its centroid. This error is also called the within-cluster variation

$$e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

The square-error for the entire clustering space containing K cluster is the sum of the within-cluster variations

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

The basic steps of the K-mean algorithm are:

- Select an initial partition with K clusters containing randomly chosen sample, and compute the centroids of the clusters.
- Generate a new partition by assigning each sample to the closest cluster center.
- Compute new cluster centre as the centroids of the clusters.
- Repeat steps 2 and 3 until optimum value of the criterion function is found or until the cluster membership stabilizes.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

learning techniques

a) **Clustering of the web users based on the user navigation patterns**

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the necessity of intelligent marketing strategies and relationship management. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining can be used for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on.

The clustering of navigation patterns presents the important concepts of Web mining and its various practical applications. The significance of this study would be to analyze the web mining approach to gather information from various web logs and to analyze the usage of a website, the total no of visits from various geographical distributed locations and thus analyze the kind of usage of a website, which will benefit the web usage and in the development of the website and its organization.

The objectives of this paper are to study and analyze metadata generated by web servers. To design a web usage mining technique for evaluating a website navigability. To generate clusters that can measure navigability of a website by identifying various characteristics by analyzing weblogs that reveal visitors. Access patterns and navigation behavior.

Related work

In this section the exploration of the web mining operations used by different operations are done here. The web usage mining process is divided into three phases. Pre-Processing, Pattern discovery and Pattern analysis.

G Langhnoja[1] has represented algorithms based on Preprocessing which include Data cleansing, User identification and user session identification. The algorithm for data cleansing requires an webserver log file as an input file which fetches the output as a log database. The algorithm for user identification requires a processed web log file which produces the number of distinct users which checks on the basis of ip address. The algorithm for session identification of a web log file consists of output of Number of sessions based on the no of user sessions**.**

AjithAbraham and [2] has represented ant colonies behavior for knowledge retrieval and management and also decision support systems science as it provides models of distributed adaptive organization which are useful to solve difficult optimization classification and distributed control problems.

K Poongathaii[3] has efficient usage of web mining which proceeds in the direction of building a efficient web usage knowledge discovery system, which gets the web user profiles at the web server, application server and core application level. He proposed the usage mining framework with Fuzzy C means clustering algorithm. The Experimentation conducted with CFuzzy means and Expected Maximization clusters in the webert data set from UCI and shows 5% to 8% better performance than CFuzzy means in terms of cluster number.

Ashish Kathuria[4] findings shows that more than 75 percent of web queries (clustered into eight classifications) are informational in nature, with about 12 percent each for navigational and transactional. Results also show that web queries fall into eight clusters, six primarily informational, and one each of primarily transactional and navigational.

TingZhong Wang[5] in his research work Web log mining has been successfully applied to a personalized recommendation system improvement and business intelligence. K-means clustering method is a common division-based clustering method, also known as K-means method, is a widely used algorithm.

Heidar Mamosian[6] Purposed systems for this problem work based on this idea that if a large number of web users request specific pages of a website on a given session, it can be concluded that these pages are satisfying similar information needs, and therefore they are conceptually related. In this study, a new clustering approach is introduced that employs logical path storing of a website pages as another parameter which is regarded as a similarity parameter and conceptual relation between web pages.

Ramya C[7] In her paper, we proposed ART1(adaptive resonance theory) neural network clustering algorithm to group users according to their Web access patterns. They compare the quality of clustering of ART1 based clustering technique with that of the K-Means and SOM(self organizing map) clustering algorithms in terms of inter-cluster and intra-cluster distances**.**

Olivia R. Liu Sheng[8] proposed a systematic Website navigability evaluation method built on Web mining techniques. To complement the subjective self-reported metrics commonly used by previous research, they developed three objective metrics for measuring Web site navigability on the basis of the Law of Surfing.

Osmar R. Zaiane[9] In his paper, discussed some data mining and machine learning techniques that could be used to enhance web-based learning environments for the educator to better evaluate the leaning process, as well as for the learners to help them in their learning Endeavour.

Pawan Lingras[10] in his paper presents clustering using a fuzzy c-means algorithm, on secondary data consisting of access logs from the World Wide Web. This type of analysis is called web usage mining, which involves applying data mining techniques to discover usage patterns from web data. The fuzzy c-means clustering was applied to the web visitors to three educational websites. The analysis shows the ability of the fuzzy c-means clustering to distinguish different user characteristics of these sites.

Conclusion

The main objective of the web usage mining is to provide information about online users aswell as online customers. The main application is to know about documentation about the online customer/user behavior, optimization or usage of a company's/organizations web portal and analysis of buying behavior and personalization of content and the layout of the websites. Web log mining has been successfully applied to a personalized recommendation system improvement and business intelligence. K-means clustering method is a common division-based clustering method, also known as K-means method, is a widely used algorithm.

References

1. *Shaily G.Langhnoja, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery" International Journal of Data Mining Techniques and Applications ISSN: 2278-2419*
2. *Ajith Abraham " Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming*"
3. *K.Poongothai "Efficient Web Usage Mining with Clustering*" *International Journal of Computer Science. www.IJCSI.org 203*
4. *Ashish Kathuria "Classifying the user intent of web queries using k-means clustering" www.emeraldinsight.com/reprints*
5. *TingZhong Wang "The Development of Web Log Mining Based on Improve-K-Means Clustering Analysis"*
6. *Heidar Mamosian "A New Clustering Approach based on Page's Path Similarity for Navigation Patterns Mining"*
7. *Ramya C "Dynamic Grouping of Web Users Based on Their Web Access Patterns using ART1 Neural Network Clustering Algorithm*
8. *Olivia R. Liu Sheng "Web mining based objective metrics for measuring website navigability*
9. *Osmar R. Zaiane "Web Usage Mining for a Better Web-Based Learning Environment"*
10. *Pawan Lingaras "Fuzzy C-Means Clustering of Web Users for Educational Sites"*
11. *"Web Mining-Concepts, Applications & Research" Jaideep Srivastava, Prasanna Desikan, Vipin Kumar*
12. *Web mining with relational clustering techniques"T.A.Runkler a, J.C. Bezdek*
13. *A Survey in Web Page Clustering Techniques"Antonio LaTorre, José M. Pena, Victor Robles, Maria S. Perez*