Study of Cluster Methods for Web Data Mining

Prof. Dr. S.P.Rasal

Abstract: Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.

From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept Conventional Information retrieval systems return long lists of ranked documents that users are forced to go through to find relevant documents. The Web search engines coupled with the ranked list presentation make it hard for users to find the information they are looking for. Clustering is the process of grouping objects together in such a way that the objects belonging to the same group are similar and those belonging to different groups are dissimilar. Clustering methods can be used in many applications. One of these application types is Web clustering where different types of objects can be clustered into different groups for various purposes. This paper deals with the different aspects of Web data mining and provides an overview about the various techniques used in this field.

Keywords: Data Mining, Clustering, Web usage mining, Web usage clustering

I. INTRODUCTION

The expansion of the World Wide Web has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized that they can be accessed efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified to better suit the demands of the Web. New approaches should be used better fitting to the properties of Web data.

Prof. Dr Subhash P. Rasal is working as HOD in Department of Electronics in Mudhoji College, Phaltan (Maharashtra) Email:subhashprasal@gmail.com)

Furthermore, not only data mining algorithms, but also information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area. Web mining involves a wide range of applications that aim 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 for example for predictive Web caching. 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.

Content Mining

The web content is the real data the web page was designed to convey to the users. It consists of several types of data such as unstructured text, graphics, sound, video and semi-structured hypertext. Content mining can be referred to as the application of data mining algorithms to the content of the web. A conceptual schema can be created that can describe the semantics of a large volume of unstructured web data to manage them. Various categories of the web content mining such as text mining which is mining of unstructured texts and multimedia data mining which is mining of multiple types of data such as unstructured and image data.

Structure Mining

The purpose of web structure mining is to classify web pages based on their organization. The structure mining can be used to categorize web pages. The potential use of this category of web mining is to generate information such as similarity between different websites .

Usage Mining Analysis of web visitors usage habits can give important clues about current market trends and help organizations to predict the future trends of potential customers. Analysis of long visit-paths of users may indicate the need of restructuring of the website to help visitors reach desired information quickly. Also, the mined knowledge can be used to offer preferred web content to visitors.

II. WEB ACCESS LOGS AND WEB USAGE MINING

The web access log is the main resource for web usage mining because it stores data pertaining to accesses of the website. The usage data can be stored in Common Log Format (CLF) or Extended Log Format (ELF). The web access log in CLF format has information of the IP address of a visitor's machine, the user id of visitor if

available and date/time of the page request. The method is a means of page request. It can be GET, PUT, POST or HEAD. The URL is the page that is requested. The protocol is the means of communication used, HTTP/1.0 for example. The status is the completion code.

2.1 WEB USAGE MINING

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation, etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. Thus the goal of each owner of a portal is to make his site more attractive for the user. For this reason the response time of each single site have to be kept below 2s. Moreover some extras have to be provided such as supplying dynamic content or links or recommending pages for the user that are possible of interest of the given user. Clustering of the user activities stored in different types of log files is a key issue in the Web community.

There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Web usage mining consists of three main steps:
(i) preprocessing, (ii) pattern discovery and (iii) pattern analysis.
Figure 1 shows the block diagram of the process of Web usage mining.

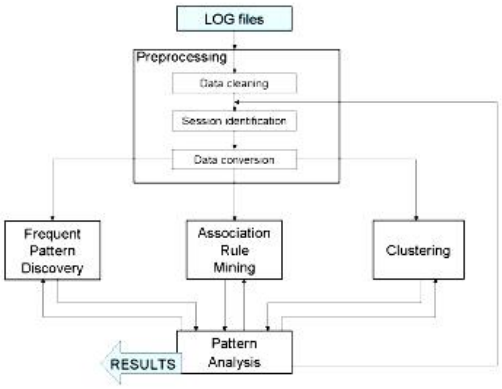

Figure 1. Process of Web usage mining

In the *preprocessing phase* the data have to be collected from the different places it is stored (client side, server side, proxy servers). After identifying the users, the click-streams of each user has to be split into sessions. It can be frequent pattern mining, association rule mining or clustering. In this paper we are dealing only with the task of clustering web usage log. In web usage mining there are two types of clusters to be discovered: usage clusters and page clusters. The aim of clustering users is to establish groups of users having similar browsing behavior. The users can be clustered based on several information. In the one hand, the user can be requested filling out a form regarding their interests, the clustering of the users can be accomplished based on the forms. On the other hand, the clustering can be made based on the information gained from the log data collected during the user was navigating through the portal. Different types of user data can be collected using these methods, for example (i) characteristics of the user (ii) preferences and interests of the user, (iii) user's behavior pattern. The aim of clustering web pages is to have groups of pages that have similar content. This information can be useful for search engines or for applications that create dynamic index pages. The last step of the whole web usage mining process is to analyze the patterns found during the pattern discovery step. The irrelevant patterns have to be filtered out, and the resulted patterns or clusters have to be validated. Some visualization techniques can help this process for the user.

III CLUSTERING ALGORITHMS

Clustering is a process of discovering groups of objects such that the objects belonging to the same group are similar in a certain manner, and the objects belonging to different groups are dissimilar. The main problems one faces when creating a clustering algorithm are the following: The objects can have hundreds of attributes that have to be taken into consideration for clustering.

One of the key issues is how to reduce this number to achieve an efficient algorithm. The type of the attributes can be diverse, and not only numerical attributes has to be handled. [2] Because of the first two problem defining a similarity function between the objects is not a trivial task. Many features and many types of attributes has to be handled efficiently. The main feature of clustering in a data mining application is the high number of objects that have to be clustered. Thus the processing time or the memory requirement of the algorithm can be huge, that has to be reduced using some heuristics. Validating the resulting clusters is also a hard task. In case of low dimensionality, when the clusters can be represented visually, the validation can be made by a human, but in case having large number of objects with high dimensionality statistical methods have to be used and indices have to be defined which can be computationally expensive. There are many algorithms in the literature that deal with the problem of clustering large number of objects. The different algorithms can be classified regarding different aspects. One of the key issue, which also determine another features of the algorithm is the basic approach of the clustering algorithm.

The aim of the **partition-based algorithms** is to decompose the set of objects into a set of disjoint clusters where the number of the resulting clusters is predefined by the user. The algorithm uses an iterative method, and based on a distance measure it updates the cluster of each object. It is done until any changes can be made. The most representative partition-based clustering algorithms are the k-means and the k-mediod, and in the data mining field the CLARANS. The advantage of the partition based algorithms that they use an iterative way to create the clusters, but the drawback is, that the number of clusters have to be determined in advance and only spherical shapes can be determined as clusters

A hierarchical algorithm: This provides a hierarchical grouping of the objects. There exist two approaches, the bottom-up and the top-down approach. In case of bottom-up approach, at the beginning of the algorithm each object represents a different cluster and at the end all objects belong to the same cluster. In case of top-down method at the start of the algorithm all objects belong to the same cluster which is split, until each object constitute a different cluster. The steps of the algorithms can be represented using a dendrogram. The resulting clusters are determined by cutting the dendrogram by a certain level. A key aspect in these kind of algorithms is the definition of the distance measurements between the objects and between the clusters. Many definitions can be used to measure distance between the objects, for example Eucledian, City-Block, Minkowski and so on. Between the clusters one can determine the distance as the distance of the two nearest objects in the two cluster, or as the two farest or as the distance between the mediods of the clusters. The drawback of the hierarchical algorithm is that after an object is assigned to a given cluster it cannot be modified later. Furthermore, like in partition-based case, also only spherical clusters can be obtained. The advantage of the hierarchical algorithms is that the validation indices (correlation, inconsistency measure), which can be defined on the clusters, can be used for determining the number of the clusters.

Density-based algorithms start by searching for core objects, and they are growing the clusters based on these cores and by searching for objects that are in a neighborhood within a radius [2] of a given object. The advantage of these types of algorithms is that they can detect arbitrary form of clusters and it can filter out the noise. DBSCAN and OPTICS are density-based algorithms.

Grid-based algorithms The grid-based algorithms use a hierarchical grid structure to decompose the object space into finite number of cells. For each cell statistical information is stored about the objects and the clustering is achieved on these cells. The advantage of this approach is the fast processing time that is in general independent of the number of data objects. Grid-based algorithms are STING, CLIQUE and Wawe Cluster.

Model-based algorithms use different distribution models for the clusters which should be verified during the clustering algorithm. A model-based clustering method is MCLUST.

Fuzzy algorithms suppose that no hard clusters exist on the set of objects, but one object can be assigned to more than one cluster. The best known fuzzy clustering algorithm is FCM (Fuzzy CMEANS).

IV CLASSIFYING THE DIFFERENT WEB CLUSTERING ALGORITHMS

There exists a great variety of Web usage clustering algorithms that can be categorized regarding several aspects. In this paper the aspects for classifying the different algorithms are the following: (i) The type of the objects to be clustered, (ii) The purpose of clustering, (iii) The clustering algorithm used, (iv) The

type of the clusters discovered (cluster overlap handling) and (v) Similarity measure. In the rest of the paper the different aspects of classifying is described in detail, and the most important and the best known web usage clustering algorithms are categorized based on the proposed aspects.

4.1 OBJECT TYPE

One aspect of categorizing a web usage clustering algorithm is the type of the objects to be clustered. As mentioned in the previous section the target of the clustering can be various.

Web pages: the most common task is to cluster web pages based on the navigation behavior of the users.

Web page sequences: the basis of the clustering algorithm can be not only the frequency of visiting a web page, but also the frequency of visiting a sequence of web pages. In this case the order of the pages also plays an important role.

User rating results: if the users have the possibility rating the different documents and web pages, then the pages can be clustered based on this information. This type of clustering is easier than clustering from log data because in this case the questions can be set up in such a way that the resulting answer vector suits the information searched for.

4.2 CLUSTERING PURPOSE

It plays a key role what the purpose of the clustering algorithm is. There exist many of purposes from which we just mention some frequently used. The most frequently used goal is to make a dynamic portal with page recommendation. For this reason the pages, documents or even user rating results can be clustered. Another important aspect is to have a portal with personalized pages or with user profiles. For this reason a user model have to be created regarding their navigational behavior. Page categorizing or page indexing makes the navigation the users easier because in this case pages that are similar in a given manner are enumerated near to each other.

4.3 CLUSTERING ALGORITHM

Because each clustering method performs differently for different purposes, the web usage clustering algorithms uses different basic clustering algorithms regarding the tasks they are accomplishing. The different basic algorithms that can be used are described in Section 3. Moreover, beside the basic algorithms, some new approaches are used in Web usage mining as well. Markov models and Fuzzy clustering algorithms are frequently used in this field. Semantic Latence Analysis (LSA) is another approach that can be used for Web usage mining.

4.4 CLUSTER OVERLAPS HANDLING

It is an interesting question how the boundary of a cluster is defined. In several cases finding hard clusters are the objective of the clustering algorithms. In this cases one object belongs to only one cluster. In other cases, however, it is enabled to have objects that belong to more than one cluster at the same time. In this case the algorithm discovers overlapping or fuzzy clusters.

4.5 SIMILARITY MEASURE

Because of the non-numerical feature of the web usage clustering problem, it is an important aspect how the similarity of the objects to be clustered is defined. In some cases the non numerical attributes are omitted, or transformed to numerical values, and Eucledian or Minkowski distance is used. In other cases new metrics are introduced in order to get a better clustering result.

V CONCLUSION

This paper deals with the problem of discovering hidden information from large amount of log data, namely, with Web usage mining. The focus of this paper is clustering among the different mining processes. After describing the task of clustering, the most common clustering methods were enumerated based on their fundamental approach.

These algorithms serve as basis for the web usage clustering that was described in detail. The different aspects of classifying the web usage clustering algorithms was described and a classification based on these aspects was provided as well.

REFERENCES

[1] BECHER, J., BERKHIN, P., and FREEMAN, E. 2000. Automating exploratory data analysis for efficient data mining. In Proceedings of the 6th ACM SIGKDD, 424-429, Boston ,MA.

[2] Kosala and Blockeel, "Web mining research: A survey," *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on KnowledgeDiscovery and Data Mining, ACM*, vol. 2, 2000.

[3] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data mining and the web.

[4] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage

patterns from web data. SIGKDD Explorations, 1(2):12{23, 2000.
[5] K.W. Tan, H. Han, and R. Elmasri. Web data cleansing and preparation for ontology extraction using WordNet. In First International Conference on Web Information Systems Engineering (WISE'00
[6] R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD, 2(1):1{15,2000.
[7] F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. In ACM SigWeb Letters, 8(3): 13-19, 1999.
[8] P. Batista, M. ario, and J. Silva, "Mining web accesslogs of an on-line newspaper," 2002.
[9] J. Hou and Y. Zhang, "Effectively finding relevantweb pages from linkage information.," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 940–951, 2003.
[10] H. Han and R. Elmasri, "Learning rules for conceptual structure on the web," *J. Intell. Inf. Syst.*, vol. 22, no. 3, pp. 237–256, 2004.
[11] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM Trans. Inter. Tech.*, vol. 3, no. 1, pp. 1–27, 2003.
[12] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in *PADKK '00: Proceedings of the 4th Pacific-AsiaConference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, (London, UK), pp. 396–407, Springer-Verlag, 2000.