Web Clustering techniques- a Study

Justina G. Nadar

Abstract— Web clustering is one of the most important research areas today because of the large amount of information available on the web. Due to the vast amount of information available on the web, a technique to classify the relevant data is a necessity. Clustering is one such data pre-processing technique. Web mining is the finding of the web resources or web users. Clustering is the process of grouping based on some common properties. In this paper we study different clustering techniques, advantages and their limitation.

Index Terms— Clustering, data pre-processing, clustering techniques, personalization, recommendation, web mining, web users.

1 INTRODUCTION

The World Wide Web has become increasingly important as a medium for information as well as for commerce. Web mining, which is the application of data mining techniques to extract knowledge from Web content, structure, and usage, is the collection of technologies to full fill this potential. Interest in Web mining has grown rapidly in its short history, both in the research and practitioner communities. The explosive growth of the Web and the increased number of users have led more and more organizations to put their information on the Web and provide sophisticated Web based services such as distance education, on-line shopping etc. However, the continuous growth in the size and use of the Internet is increasing the difficulties in managing the information. Thus, an urgent need exists for developing new techniques in order to improve the Web performance.

Clustering is grouping of objects with common properties in the same group and different objects are placed on different groups based on similarity comparison. Web user clustering is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters. Clustering of web users can be applied to various applications such as personalization and recommendation of E-commerce websites. Personalization and recommendations helps to understand customer preferences on time.

The paper presents a detail survey of various clustering techniques their advantages and limitations. Literature survey is cited in section 2. Mining techniques are discussed in section 3. Clustering techniques are explained and compared in section 4 and 5. Section 6 concludes the paper.

2 LITERATURE SURVEY

In this section we site the relevant past literature that use different strategies to cluster web user's most common methods are machine learning and statistical methods.

Clustering by constructing page hierarchies and attribute oriented approach are introduced by Yongjian Fu [1]. Using the access patterns from log files of the users a page hierarchy can be constructed and a generalization technique called attribute oriented induction is applied to generalise the pages. The author thinks it is an efficient to cluster the user data using the proposed approach.

Yanchu Zang[2] proposed a model, Latent semantic model to capture semantic associations among user transactions. Web transaction data between web visitors and web functionalities usually convey users' task-oriented behaviour patterns. Clustering web transactions, thus, may capture such informative knowledge and build user profiles. Profiles are associated with different navigational patterns.

In the paper Time Aware Web Users Clustering [3], the author introduced a new approach. Web users are clustered based on time locality. The paper focuses on two aspects: different users page preferences and time dependencies involved in the usage of navigational patterns. It emphasizes the need to discover similarities in users

- *Justina G. Nadar currently pursuing masters degree in Information Technology in Mumbai University, India,E-mail:justina.sharon@gmail.com* accessing behaviour with respect to the time locality of their navigational act.

Yunjuan Xie[4] uses the Dempster-Shafers theory of combining evidence to find and group pages that are frequently visited into different classes. Clustering task is added to capture the uncertainty among the user behaviour.

Cyrus Shahabi and Amir M. Zarkesh[5] together speaks about knowledge discovery in web navigation. A method is described to detect the navigation path accurately. Path clustering method based on the similarity of the history of user navigation is proposed. This approach is capable of capturing the interests of the user which could persist through several subsequent hyper-text link selections.

One of the researchers Christos Bouras[6] used WordNet based approach to cluster the user interests. The sessions obtained from user's navigation path are enriched using WordNet hypernyms and the hypernyms are clustered to provide more generalised keywords for clustering user data.

3 MINING TECHNIQUES

Web mining techniques can be classified as Web content mining, Web structure mining and Web usage mining. Web content mining describes the automatic search of information resource available online, and involves mining web data contents. Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites. Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behaviour while the user interacts with Web.

4 TYPES OF CLUSTERING TECHNIQUES

It is difficult to provide a crisp categorization of clustering methods because these categories may overlap so that method may have features from several categories.

4.1 Partitioning methods

It is a very popular method of clustering due to its simple approach. A partitioning method constructs k partitions of the data, where each partition represents a cluster. It divides the data into k groups such that each group must contain at least one object. Partitioning methods conduct one-level partitioning on data sets. Partitioning methods adopt exclusive cluster separation such that each object should belong to exactly one group. Partitioning methods are generally distance based. It uses iterative relocation technique that improves the clustering by moving objects from one cluster to another. [12]

1) Kmeans: It is based on a centroid based iterative technique. The centroid of a cluster is its center point. It first randomly selects cluster center. It computes new mean and the objects are reassigned to the new cluster in each iteration.

2) Bisecting Kmeans: It is a variant of K-means. Bisecting Kmeans [7] splits a single cluster into two sub clusters at each bisecting step. Bisecting Kmeans is more efficient is the no of clusters is large. The computation time is reduced as only the data points of one cluster and two centroids are involved in the computation.

4.2 Hierarchical methods

A hierarchical method creates a hierarchical decomposition of a given set of data objects. Hierarchical clustering can be distance based or density based. It is widely used method and generally combined with various other approaches. [12]

1) AGNES: An agglomerative hierarchical clustering method uses a bottom-up strategy. It starts with each object form its own cluster and iteratively merges clusters to form larger cluster until a termination condition is met.
2) DIANA: A divisive hierarchical clustering uses top-down strategy. It starts by placing the objects in a single cluster which is the root. It then divides the cluster into smaller sub-clusters.
3) BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies is used to cluster large numeric data. BIRCH uses the notion of clustering feature to summarize a cluster and clustering feature tree to represent a cluster hierarchy. Summarizing a cluster using clustering feature can avoid storing information thus it is efficient in space. It clusters the leaf nodes and removes sparse clusters as outliers and group dense clusters into large ones. See T.

Zhang, R. Ramakrishnan [8] for detailed discussion on BIRCH.

4.3 Density-based methods

A density based clustering is to continue growing a given cluster as long as the density in the neighbourhood exceeds some threshold.

1) DBSCAN: Density Based Spatial Clustering with Noise first finds the objects that have dense neighbourhood and connects those objects and their neighbourhood to form clusters. [12]

4.4 Grid-based methods

Grid based methods quantize the object space into finite number of cells that form a grid structure. It has fast processing time which is dependent on the number of cells rather on the data objects. [12]

4.5 Probability based cluster models

Probability based cluster models, cluster the objects based on some defined threshold. Probability based cluster models are more effective and general to be used. Probability based cluster models are often used to measure the expected outcome to the real outcome. It is used in cluster analysis to find hidden categories. [12]

4.6 Using vector space model

The tdf-idf weighting scheme is used in the vector space model with cosine similarity to determine the similarity between the two documents or texts represented in the vector space. If a term frequently appears in a document then the document containing the term should be retrieved it is known as term document frequency. The terms scarcity in a document is known as inverse term frequency.

In vector space model weighting is very formal but is shown effective in various experiments [9]. It is used as a tool in scoring and ranking document's relevance to a given user query.

4.7 Using bayesian models

Bayesian model is a probabilistic classifier. It applies the Bayesian theorem with assumption on the features. Bayesian probability theory can be used represent degrees of belief in uncertain propositions. It applies prior probability before observing any information and posterior probability after observing. A Bayesian model allows defining predictive distributions and can be combined with traditional hierarchical method to get good quality clusters. John W. Law and Peter J. Green [10] as discussed in paper [10].

4.8 Using WordNet

WordNet is an online lexical database available for English language. It groups the English words into sets of synonyms called synsets. WordNet also provides a short meaning of each synset and semantic relation between each synset. WordNet also serves as a thesaurus and an online dictionary which is used by many systems for determining relationship between words. Thesaurus is reference work that contains a list of words grouped together according to the similarity of meaning. Semantic relations between the words are represented by synonyms sets, hyponym, and metonym trees. WordNet are used for building lexical chains according to these relations [11].

Paper [6] discuss about use of WordNet hypernym in clustering. Using WordNet in clustering enhances the results to more refined form. It uses the external knowledge extracted from the WordNet database to enhance the bag of words prior to clustering.

5 SUMMARY TABLE OF CLUSTERING TECHNIQUES

In this section, we present a comparison in a tabular format about the various clustering techniques their types, sub-types, concept, advantages, disadvantages and applications.

TABLE 1. SUMMARY TABLE OF CLUSTERING TECHNIQUES

Methods	**Sub-types**	**Concept**	**Advantages**	**Disadvantages**	**Applications/ work done**
Partitioning Methods	K-means [12]	A centroid based iterative technique. It computes new mean and the objects are reassigned to the new cluster in each iteration.	Efficient with small and medium sized data.	Gives only spherical shaped clusters	Recommender system
	Bi-secting Kmeans [7]	Splits a single cluster into two sub clusters at each bisecting step.	Good quality clusters and high performance	Not suitable for text documents.	Recommender Systems
Hierarchical Methods	AGNES [12]	Each object is assigned a cluster of its own. Similarity between the clusters is determined and the clusters are merged.	High performance	Not scalable to large datasets	Generalization
	DIANA [12]	All objects are in the same cluster initially and then split according to some principle.	High Performance	Cannot undo what was done previously	Generalization

	BIRCH [8]	It uses a clustering feature tree to represent a cluster hierarchy	Space efficient	Low performance	Generalization
Density Based Methods	DBSCAN [12]	It connects core objects and their neighborhoods.	High processing speed	Cannot handle high dimensionality data	To find Outliers
Grid Based Methods	STING [12]	Grid based methods quantize the object space into finite number of cells that form a grid structure.	Query independent	Costly and less accurate	In databases
Probability Based Cluster Model	Probability model [12]	Probability based cluster models, cluster the objects based on some defined threshold.	Efficient	Cannot handle uncertainties	E-commerce websites
Using Vector Space Model	VSM [9]	It uses tdf-idf weighting scheme.	High Efficient	Memory consuming	Document classification
Using Bayesian Model	Bayesian Model [10]	Bayesian model is a probabilistic classifier. It applies the Bayesian theorem with assumption on the features.	Efficient	Cannot handle uncertainties	Document classification

Using WordNet	W-kmeans [6]	Use an online thesaurus to find similarities between words or construct a word hierarchy	Highly Efficient	Not Scalable	Personalization engine

6 CONCLUSION

As clustering improves, more efficient information can be retrieved from the web about the users behaviour and access patterns, this can be applied to classify data, find loyal customers, to provide online support and in E-commerce websites.

Due to the rapid growth of the internet the information is overloaded. This problem can be solved by using clustering techniques to classify the data to help the users. We described different types of web clustering methods. Here there is a need to develop a system efficient of clustering the data. Clustering using partitioning and hierarchical methods are simple to implement and more efficient. Clustering can also be combined with generalisation techniques to get more refined results.

ACKNOWLEDGMENT

I would like to express my thanks to Lord Almighty; my parents for their support. Sincere thanks to my project guide for reviewing my work. I am thankful to my Institution and my University.

REFERENCES

[1] Y. Fu, K. Sandhu and M. Shih, "Clustering of Web Users Based on Access Patterns", Computer Science Department University of Missouri Rolla, 2001.

[2] Y. Zang and G. Xu," Latent usage approach for clustering web transaction and building user profile." School of Computer Science Victoria University, Australia, 2005.

[3] G. Sophia, A. Vassiliki and I, "Vakali. Time aware web users clustering." Senior member, IEEE.

[4] Y. Xie and V. Vir, "Web users clustering from access log using belief function." Computer Science Department Louisiana Tech University, Canada, Oct 22-23 2001.

[5] C. Shahabi, A. Zarkesh, J. Adibi and V. Shah, "Knowledge Discovery from Users Web-Page Navigation." Computer Science Department, University of California.

[6] Christos Bouras, Vassilis Tsogkas. Clustering user preferences based on W-kmeans, 2011, Seventh International Conference IEEE.

[7] Yanjun Li, Soon M. Chung. Parallel Bisecting K-means, Wright State University USA, 2007.

[8] T. Zhang, R. Ramakrishnan and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In proc. 1996 Int. Conf. Management of data, Canada June 1996.

[9] Dik L. Lee, Huei Chuang, Kent Seamons, "Document Ranking and Vector Space Model", Hong Kong University of Science and technology, March 1997.

[10] John W. Law, Peter J. Green, "Bayesian Model Based Clustering Procedure", University of Bristol, UK, January 2007.

[11] Mahlon Lovett, "http://wordnet.princeton.edu/" Office of Communications, Princeton University, August 2014.

[12] Lior R., Oded M."Clustering Methods" Data Mining and Knowledge Discovery Handbook, Tel Aviv University.