ISSN 2320 – 2602
International Journal of Advances in Computer Science and Technology (IJACST), Vol. 3 No.2, Pages : 07 – 11 (2014)
Special Issue of ICCSIE 2014 - Held during February 16, 2014,Bangalore, India

An efficient cloud based framework for Web recommendation systems

Sowmya.K.Menon[1], Varghese Paul[2], M.Sudheep Elayidom[3] Ratan Kumar[4]
[1]Assistant Professor, SNGIST, Manjali, Kerala, India, sowmyamenon@rediffmail.com
[2]Associate Professor, CUSAT, Kochi, India,vp.itcusat@gmail.com
[3]Associate Professor, CUSAT, Kochi, India, sudheepelayidom@hotmail.com
[4]Ratan Kumar, software engineer, tracxn.com, Bangalore, ratancs@live.com

Abstract*:* In this era of online and social network revolution, to have an efficient system which recommends good web sites of interest to users is a need of the time. The proposed system accomplishes this by analyzing the user's browsing habits. The system is based on web usage mining concepts, which analyses the user's web usage statistics and data mining tasks like clustering. Here the browsing habits of a group of users are analyzed and the users are grouped to different clusters in the server such that the users in the same cluster have similar browsing habits to visit similar web sites. The core idea is the system will recommend the frequent web sites visited by other users in the same cluster to a typical user of a cluster. The entire system is implemented over a cloud framework of many users and a cloud server based on the Google App Engine.

Key words: Web usage mining, cloud based framework, Google App Engine, Web recommendation, clustering

INTRODUCTION

A web based recommendation system is a web based interactive software agent. In web based recommendation system, the system will predict the user's web browsing behavior by usage data analysis. The proposed web recommendation system consists of two modules; Client side modules and server side modules. The client-side module prepares user browsing data while the server side modules clusters and prepares recommendations for users.

One of the earliest and widely used technologies for building recommender systems is Collaborative Filtering [2, 6]. Collaborative filtering approach build a model from a user's past behavior such as those items previously or numerical ratings given to those items as well as similar decisions made by other users, then use that model to predict items (or ratings for items) that the user may have an interest in. A collaborative filtering algorithm searches a large group of people and finding a smaller set with tastes similar to a particular user. To create a ranked list of suggestions, it looks at other things they like and combines them.

A product is recommended to the current user if it is highly rated by other users who have similar interests. A web based recommendation system can also be based on a hybrid approach that combines item based and collaborative filtering approaches [1]. In this paper we investigate a cloud based framework for web based recommendation system. In recent years there has been increasing interest in applying web usage mining techniques to build web recommender systems. In Web usage recommender systems, either web server logs or direct client browsing habits are given as inputs, and make use of data mining techniques such as association rules and clustering to extract navigational patterns, which are then used to provide recommendations. User browsing history is recorded in web server logs, which contains much hidden information regarding users and their navigation. They can be used for user rating or feedback in deriving user models. In conventional recommender systems, browsing patterns are generally derived as off-line and online processes [3]. Association rules is one of the most commonly used approach for web usage recommender systems. But in this work, clustering is used as a main process for recommendation systems.

Web usage mining has lots of future prospects. Reasons are very simple: With the web revolution, the way companies are doing businesses has been changed. Online business, mainly characterized by electronic transactions through Internet, has provided effective way of doing business. Amazon.com has now become an "on-line Wal-Mart". Web is nothing more than a place where transactions take place to most companies unfortunately. Business people have to realize that millions of visitors interact daily with Web sites around the world; huge amounts of data are being generated, nowadays known by the buzz word "Big data". They also have to realize that this information is very precious to the company in the fields of understanding customer behavior and other CRM related activities [8, 10]. But in this paper, rather than a business perspective, a customer centric view is adopted. For ordinary users they would like to know the web sites that may interest them. This system recommends the most interesting web sites for them. Hence this system proposes another application area for web usage mining rather than the conventional e-commerce related benefits.

ISSN 2320 – 2602
International Journal of Advances in Computer Science and Technology (IJACST), Vol. 3 No.2, Pages : 07 – 11 (2014)
Special Issue of ICCSIE 2014 - Held during February 16, 2014,Bangalore, India

PROBLEM STATEMENT

To propose a web mining methodology based on cloud computing, to recommend most suited web pages that may interest a particular user. The problem is to identify those types of web sites that are more prone to be used by the users.

There are many applications for this type of information, such as making recommendations for online shopping, suggesting interesting web sites, or helping people find music and movies. The end product of this work is an add-on which will filter the history of browser and all the searches that user make. This will help to clearly categorize the searches made by user into different classes. On the basis of this classification, proper recommendation can be done. The project shows how to build a system for finding people who share tastes and for making automatic recommendations based on things that other people like.

CONCEPTS USED

The main concepts used are web mining, clustering and Cloud based framework. This paper proposes a framework for analyzing surfing behavior of internet users. Google App Engine is a cloud computing platform for developing and hosting web applications in Google-managed data centers.

Web mining is the application of data mining techniques to extract knowledge from web data. Web data mining is the application of data mining techniques to extract knowledge from Web data i.e. Web Content, Web Structure and Web Usage data.

Data clustering is a method for discovering and visualizing groups of things, people, or ideas that are all closely related. Clustering is used frequently in data-intensive applications. Using clustering technique we are grouping users having similar behavior.

All the user browsing details are to be stored in a database called Big Table in the proposed system. Google uses as data storage facility called Big table. Big table is a distributed, persistent, multidimensional sorted map. Big table is not a relational database. In Big table you can store strings under an index which consists of a row key, a column key and a timestamp. Big table is build upon the Google File System and stored in an immutable data structure called SSTable. The application can define how many entries based on the timestamp should be kept. Alternatively the application can also specify how long entries should be kept. Big table will clean the obsolete data by deleting the SSTable which contains irrelevant data.

As shown in figure 1, there are many different types of web mining. In web content mining, the web document contents are analyzed to find interesting patterns. In web structure mining, the web site navigation structures are mined to study the navigation behaviors of web pages. In web usage mining, the user's web usage statistic is analyzed. This system is a typical application of web usage mining.

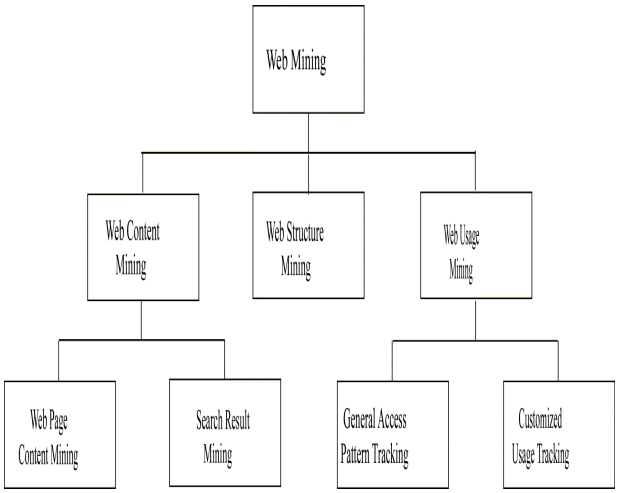

Fig 1: The web mining taxonomy

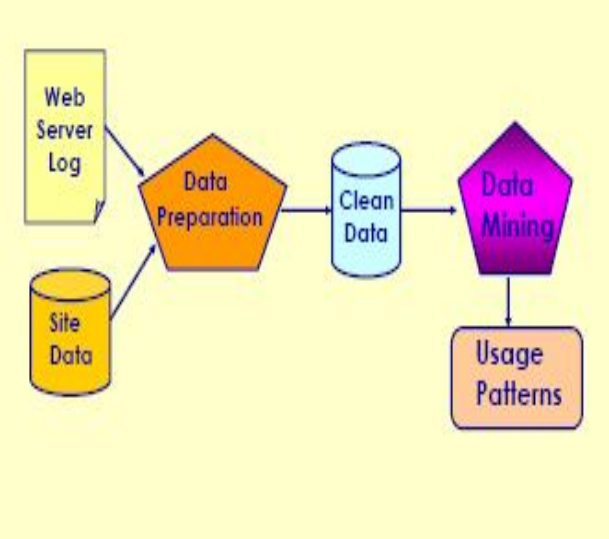

Fig 2: The Web usage mining process

The above diagram shows the various phases of web usage mining. First, the web server log is analyzed to see the usage statistics of the web site. Then one of the most important data mining process takes place namely data pre-processing, which may involve preparation and cleaning of data, which may make the data suitable for further data mining process [5, 7]. The output from the system is the usage patterns for the web sites. But in this system, rather than collecting the usage statistics from web servers and cleaning it, the actual browsing patterns are directly collected from the client machines it selves for different users browsing different web sites, since the analysis can be more realistic and dependable as explained in the coming section.

ISSN 2320 – 2602
International Journal of Advances in Computer Science and Technology (IJACST), Vol. 3 No.2, Pages : 07 – 11 (2014)
Special Issue of ICCSIE 2014 - Held during February 16, 2014,Bangalore, India

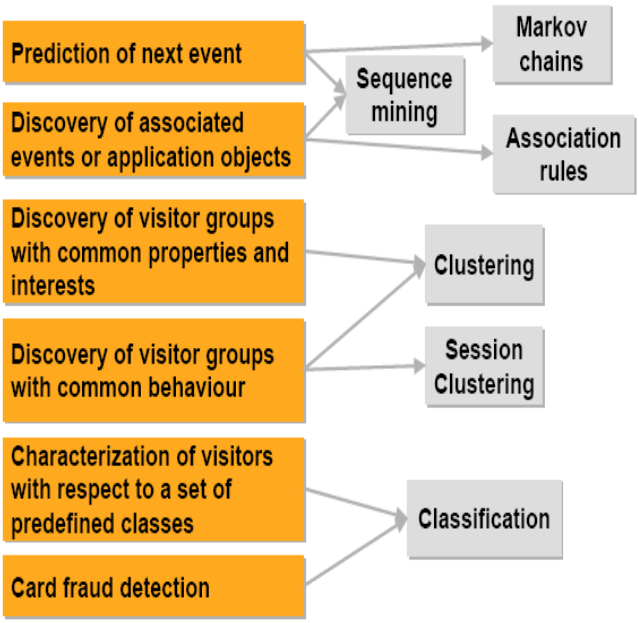

Fig 3: Types of web usage mining tasks

The above diagram shows the different types of web usage mining tasks like sequence mining, association rules mining, clustering, and classification. From the above diagram, it is clear that this work falls to the discovery of visitor groups with common behaviors using clustering.

PROPOSED ARCHITECTURE

The proposed architecture used in this system is client server architecture. Users having similar browsing behaviors are clustered and stored in the server. The clustering of similar users and preparation of recommended web sites will take place in the server.

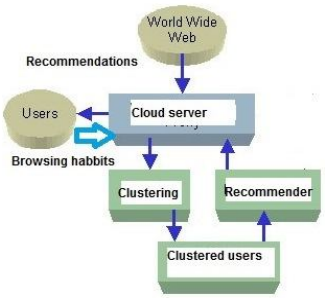

Fig 4: The proposed Architecture for the Recommendation

In the above figure, the proposed architecture for the recommendation system is shown. As the users browse websites, the client part of the software (installed as a plug-in in the user's browser), collects the browsing information and communicates to the server on a periodic bases. The server is a cloud server based on the Google app engine architecture which does the actual clustering based on the k-means algorithm. The recommended web sites used by clustered users are returned by this recommendation system to the client plug-in as shown in the above architecture diagram. These recommendations are given to the user for further browsing. Google App Engine lets you run web applications on Google's infrastructure. App Engine applications are easy to build, easy to maintain, and easy to scale as your traffic and data storage needs grow. With App Engine, there are no servers to maintain: You just upload your application, and it's ready to serve your users.

This type of algorithm is quite different from hierarchical clustering because it is told in advance how many distinct clusters to generate. The algorithm will determine the size of the clusters based on the structure of the data. K-means clustering begins with *k* randomly placed *centroids* (points in space that represent the center of the cluster), and assigns every item to the nearest one. After the assignment, the centroids are moved to the average location of all the nodes assigned to them, and the assignments are redone. This process repeats until the assignments stop changing. Figure 5 shows this process in action for five items and two clusters.

In the first frame, the two centroids (shown as dark circles) are placed randomly. Frame 2 shows that each of the items is assigned to the nearest centroid—in this case, A and B are assigned to the top centroid and C, D, and E are assigned to the bottom centroid. In the third frame, each centroid has been moved to the average location of the items that were assigned to it.

When the assignments are calculated again, it turns out that C is now closer to the top centroid, while D and E remain closest to the bottom one. Thus, the final result is reached with A, B, and C in one cluster, and D and E in the other. The function for doing K-means clustering takes the same data rows as input as does the hierarchical clustering algorithm, along with the number of clusters (k) that the caller would like returned.

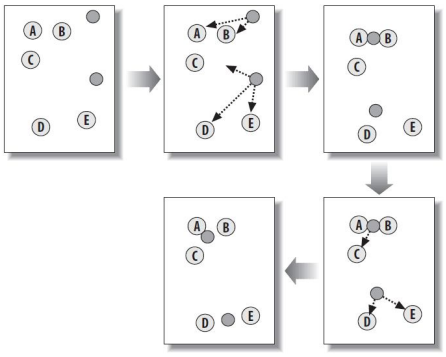

Fig 5: K-Means Clustering

Steps for the K-Means algorithm:

- Initially, the number of clusters must be known, or chosen, to be K say.
- The initial step is to choose a set of K instances as centres of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.
- Next, the algorithm considers each instance and assigns it to the cluster which is closest.
- The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments.
- This process is iterated until a stopping criterion is met, i.e., when there is no further change in the assignment of the data points.

On implementation, this system can be tested among a group of say, 100 users, by observing their browsing habits for a period of say, 1 month. They may install the plug-in from the web site and their browsing habits may be monitored for a period of two weeks. It should be proved that the users in the same cluster as returned by the software, are visiting the same web sites in next two weeks as suggested by the recommendation system.

ADVANTAGES OF PROPOSED SYSTEM

The main advantage of the proposed system is that users can get an idea about the websites that may interest them using this system. The recommendation is based on user's who have the same browsing behavior. The complete concept is based on cloud implementation which is freely available on the internet in the forms of Google App engine, Big table and so on for easy deployment.

The tool can be made available as a plug-in which can easily be downloaded and installed in any client browsers. The clustering process and generation of recommendations is fully separated from the users view and taken care by a server machine. Hence the system is easy to use.

FUTURE SCOPE

Sites like Amazon.com, Pandora radio, Netflix uses the concepts of recommendation systems. It has been in recent use, to help user get recommended according to their likes. Predictive analytics have successfully proliferated into applications to support customer recommendations, customer value and churn management, campaign optimization, and fraud detection.

This work has a future scope in finding the browsing behavior of a specific group of internet users such as customers of electronic products, books etc and giving recommendation according to it. The recommendation is based on those users who have the same browsing behavior. Thus a user can get some websites of their own interest.

The clustering algorithm that is proposed in this work is the classical K-Means algorithm. Since the field of data mining algorithms is emerging with tremendous pace, some new generation fast clustering algorithms may be developed and tested for the clustering purpose. Also instead of the Google App Engine/ Big table frame work in this work, Hadoop /Map reduce frame works may also be used and tested as a future work.

One growing area of research in the area of recommender systems is mobile recommender systems. With the increasing ubiquity of internet-accessing smart phones, it is now possible to offer personalized, context-sensitive recommendations [2, 4]. This is particularly a difficult area of research, as mobile data is more complex than conventional recommender systems, since it is heterogeneous, noisy, requiring spatial and temporal auto-correlation. One example of a mobile recommender system is one that offers potentially profitable driving routes for taxi drivers in a busy city with frequent traffic jams. The system uses machine learning techniques and reasoning process in order to adapt dynamically the mobile recommender system to the evolution of the user's interest.

ACKNOWLEDGEMENTS

We would like to thank the students of Cochin University namely, Mr. Yashwant Kumar Singh, Ms. Rashmi M.R, Mr. rahul ranjan and Mr. poorvank Bhatia who have contributed in the implementation and documentation part of this work.

CONCLUSION

In this work, a web site recommendation system for users is proposed. It does the recommendation in three phases. In the first phase, the user's browsing habits are collected on a periodic basis from the client machines to server. In the second phase, it clusters the users according to the frequent sites visited by them. This takes place on the server side. In the third phase, those web sites that are visited by the users in the same cluster are recommended to users. The main advantages of the proposed system are that by mapping user internet browsing behavior to an abstract layer of semantic features, we can not only recommend other website in the same class of websites that "match" the user model, but also understand the customer's "tastes" and recommend websites across categories.

Our approach also allows us to "explain" the recommendations in terms of qualitative features which enhances the user experience and helps build the user's confidence in the recommendations.

Web usage mining has become as an essential tool for realizing user-friendly and business-optimal Web services. Web usage mining is mainly used by e-commerce sites to organize their sites with a view to increase profits. Nowadays, it is also used by search engines to improve search quality and to evaluate search results. But in this paper an application is proposed which is beneficial for the users rather than conventional e- commerce applications, which can be very well adopted by online business organizations [9, 11].

REFERENCES

[1] Jia Li and Osmar R. Ziane, "Combining Usage, Content, and Structure Data to Improve Web Site Recommendation***", In*** *Proceedings of the 5th International Conference, EC-Web* , Spain, 2004, pp 305-315.

[2] Harita Mehta, Shveta Kundra Bhatia, Punam Bedi and V. S. Dixit, (November 2011). Collaborative Personalized Web Recommender System using Entropy based Similarity Measure. *IJCSI International Journal of Computer Science Issues*, 8(6), pp. 231-240.

[3] Toby Segaran, *Programming Collective Intelligence*,1 st ed, O'Reilly, 2007.

[4] G. Adomavicius and A. Tuzhilin,(2005).Towards the next generation of recommenender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. on Data and Knowledge Engineering*, 17(6), pp. 734-749.

[5] I.J Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Elsevier, 2011.

[6] F.Y. Tani, D.M. Farid and M.Z. Rahman(January 2012). Ensemble of Decision Tree Classifiers for Mining Web Data Streams. *International Journal of Applied Information Systems* 1(2), pp. 30-36.

[7] Ungar L.H. and Foster D.P, "Clustering Methods for Collaborative Filtering", In the Proceedings of the Workshop on Recommender Systems, AAAI Press, 1998.

[8] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl, ".Analysis of recommendation algorithms for e-commerce", In *Proc of ACM Conference on Electronic Commerce*, Minneapolis, MN, USA , 2000, pp. 158–167.

[9] Upendra Shardanand and Patti Maes, "Social information filtering: Algorithms for automating "word of mouth"", In *Proc of ACM CHI'95 Conference on Human Factors in Computing Systems*, Denver, Colorado, USA , 1995, pp. 210– 217.

[10] W. Wong and A. Fu. Incremental document clustering for web page classification, 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1 .1.21.2379&rep=rep1&type=pdf

[11] Arbee L.P, Chen Yi-Hung Wu, Yong-Chuan Chen, "Enabling personalized recommendation on the web based on user interests and behaviours", In *proc. of 11th International Workshop on research Issues in Data Engineering*, Heidelberg, 2001, pp.17-24.