THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Collaborative Filtering Approach Based on Clustering for Big Data Application

Divya S.
PG Scholar, Department of Information Technology,
Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
Kanya Rajesh R.
PG Scholar, Department of Information Technology,
Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
Rini Mary Nithila I.
PG Scholar, Department of Information Technology,
Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India
Vinothini M.
PG Scholar, Department of Information Technology,
Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India

Abstract:
As the technology is growing day by day,the amount of data created per second is more than we can imagine. The large amount of data created are also called as Big Data. In order to access those large datasets we have to go for parallel processing systems. By using the parallel processing systems, the large data are split and n no of servers are used for storing the data and parallelly executed. This paper deals with collaborative filtering method for effective searching and retrieving of data from large datasets with the help of clustering.Clustering is nothing but grouping of related elements. Collaborative filtering is a recommender system that predicts by analysing the interests of the users. Data sparsity will be reduced, accurate results will be produced and online execution time will be minimised.

Keywords: *Big data, clustering, collaborative filtering, data sparsity*

1. Introduction

Big data is a fast growing data having terabytes or petabytes of information. The data are stored in variety of files. Data classification is very important in order to access the small datas in less time and to get accurate answers. Data Accuracy is one of the major isuues while dealing with the variety of data, as the data is large it is difficult to provide accuracy. Accuracy can be achieved only by providing a format in which the data can be arranged for quick access .Big data applications will be of great success if they support self service analysis capability.

Big data can be described by the following characteristics: Volume –It defines the size of the data. Big data can be referred only by its amount of data, it is really big when compared to other data. Variety –It refers to the category of data that can be searched and analysed for the betterment of the society and organisation .Velocity –It refers to the speed and how fast the data is generated. Variability – It is the inconsistency of the data and shows how to handle and manage the data effectively. Veracity - Accuracy of analysis depends on the veracity of the original data. Complexity – Managing the data is not an easy task the datas has to arranged in an order in order for easy accessing Collaborative filtering method is used by some recommender systems, it deals with methods that can be used to process large amount of data and also provide extra knowledge about the data without users involvement. Clustering can be used to gather related elements. The main advantages of clustering is to reduce data sparsity. Clustering based collaborative filtering is used mainly to make decision within acceptable time, to explore the large volumes of data and extract useful information or knowledge for future actions,and to generate ideal recommendations from so many services.

2. Collaborative Filtering

Everyone will have something common in choices and we agree recommendations. Taking the online services we will like to buy the items bought by the users of similar tastes. Applications that are using large amount of datas can be searched easily by using the collaborative filtering . It is a method of making automatic predictions (filtering) about the interests of a user by collecting taste

information from many users (collaborating). A collaborative filtering system for television tastes could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes). Friends and relative strangers that we are connected to via Twitter, Facebook, YouTube, and other sites generally happen because of some kind of connection we have with them: work, interest area, belief, social connection. But of course those relationships are never confined to what we share in common. Folksonomy is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content; this practice is also known as collaborative tagging, social classification, social indexing, and social tagging. Social bookmarking is a method for Internet users to organize, store, manage and search for bookmarks of resources online. Unlike file sharing, the resources themselves aren't shared, merely bookmarks that reference them. The figure illustrates the searching process where the collaborating filtering approach is used to give extra information.
CF is one of a handful of learning-related tools that have had broadly visible impact: Google, TIVO, Amazon, personal radio stations. The web services like online shopping architecture will be having these architecture. Around two hundred servers will be containing the products we are searching. Each server will have a secondary server which is nothing but mirroring. The products are clustered in each server. For eg, Mobiles of all brands will be available in one server, if we search the mobiles in the application the request is directed to the particular server that contains the mobiles instead of searching all the servers. Thus by clustering the time is saved and accurate results will be produced. Collaborative filtering plays a major role by providing extra information rather than we are searching by maintaining a cache for every single user and show them as" Customers who bought this also viewed" The tastes are analysed for every user and predictions are done by the collaborative Filtering approach.

3. How Big data is processed within less amount of time?
User is responsible for generating request, the request may be the name of the product and related items. The request is handled by the load balancer and the server will process it and send that to the user. Controller will be handling the load balancers and activate only the load balancer which handles the server that contains the user request. Load balancer will be acting as an interface between the controller and the server. The user requests will be loaded to the concerned server and the processed data from the server will be provided to the user through the load balancer. The load balancer will get the IP address from the user after the request is given by the user. The concerned server will be connected then the products will be searched and the results will be produced to the user. The user requests will be directed to the concerned server by the load balancer. By collaborative clustering algorithm the user will be provided by the results based on active users interest.

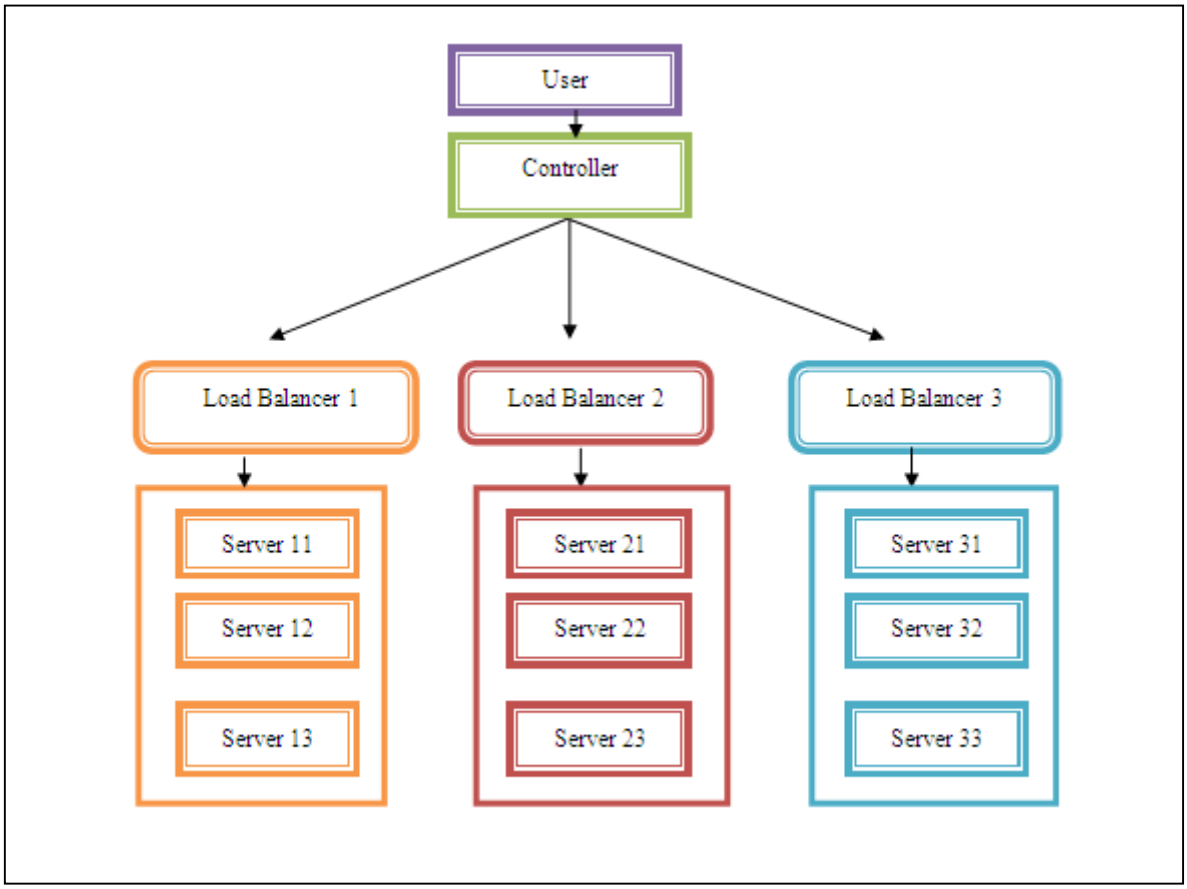

Figure 1: Clustering based collaborative Filtering

4. Collaborative Filtering Algorithm

A traditional collaborative filtering algorithm represents a customer as an *N*-dimensional vector of items, where *N* is the number of distinct catalog items. The components of the vector are positive for purchased or positively rated items and negative for negatively rated items. To compensate for best-selling items, the algorithm typically multiplies the vector components by the inverse frequency making less well-known items much more relevant.3 For almost all customers, this vector is extremely sparse. The algorithm generates recommendations based on a few customers who are most similar to the user. It can measure the similarity of two customers,A and B, in various ways; a common method is to measure the similarity is:

$$\text{Similarity(A and B)} = \frac{\text{relevant items} \cap \text{retrieved items}}{\text{relevant items}}$$

The algorithm can select recommendations from the similar customers' items using various methods as well, a common technique is to rank each item according to how many similar customers purchased it. Grey sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering. Black sheep are the opposite group whose idiosyncratic tastes make recommendations nearly impossible. Although this is a failure of the recommender system, non-electronic recommenders also have great problems in these cases, so black sheep is an acceptable failure.

5. Recommender Systems

Recommender systems are most popular now a adays because of the applications using Big data. It recommends to users based on tastes and preferences. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviors, activities or preferences and predicting what users will like based on their similarity to other users. Content based filtering is based on the description of the items, it uses some keywords.The algorithm recommend the system the user liked in the past, instead of reviewing users profile, it will focus on users interests. Knowledge based deals with the preferences that satisfies the user's need. Hybrid recommender systems uses everything. Risk aware recommender system is used to avoid the disturbances that cause the user to give bad answers.

6. Conclusion and Future Work

The online execution is reduced, since we are searching the relevant servers not the entire servers. Collaborative approach is a costly approach, the maintenance of many servers is not easy, but accurate results will be produced. This paper summarizes how we can effectively search and produce accurate results. Big data need large storage space as well as maintenance, effective storage can make the search easier. Extra knowledge can be improved by using the semantics, it is nothing but meaning based search. By this semantic similar services can be clustered together and recommendations can be given even if there are very few ratings.

7. References

i. M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.
ii. X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.
iii. A. Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.
iv. Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE BigData, pp. 403-410, October 2013.
v. J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. OSDI '04, pages 137–150, 2004
vi. Z. Liu, P. Li, Y. Zheng, et al., "Clustering to find exemplar terms for keyphrase extraction," in Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 257-266, May 2009.
vii. T. Niknam, E. Taherian Fard, N. Pourjafarian, et al., "An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering," Engineering Applications of Artificial Intelligence, vol. 24, no. 2, pp. 306-317, March 2011.
viii. http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics