A Survey on Clustering Based Collaborative Filtering Approach for Big Data Application

Shweta Bhonde[1], Mirza Baig[2]
[1]*M. Tech. Student, Department of CSE, JD College of Engineering and Management, Nagpur, India*
[2]*Professor, Department of CSE, JD College of Engineering and Management, Nagpur, India*

***Abstract*: The growth of the Internet has made it much more difficult to effectively extract useful information from all the available online information. As a result, service-relevant data become too big to be effectively processed by traditional approaches. In view of this challenge, a clustering-based collaborative filtering approach is proposed in this paper, which aims at recruiting similar services in the same clusters to recommend services collaboratively. Technically, this approach is enacted around two stages. In the first stage, the available services are divided into small-scale clusters, in logic, for further processing. At the second stage, a collaborative filtering algorithm is imposed on one of the clusters. Since the number of the services in a cluster is much less than the total number of the services available on the web, it is expected to reduce the online execution time of collaborative filtering.**

***Keywords*: Big data application, cluster, collaborative filtering, etc.**

1. Introduction

The motivation for collaborative filtering comes from the idea that people often get the best recommendations from someone with similar tastes to themselves. Collaborative filtering explores techniques for matching people with similar interests and making recommendations on this basis. Collaborative filtering algorithms often require [1] users' active participation, [2] an easy way to represent user's interests to the system, and [3] algorithms that are able to match people with similar interests. Typically, the workflow of a collaborative filtering system is

- A user expresses his or her preferences by rating items (e.g. books, movies or CDs) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain.
- The system matches this users ratings against other users and finds the people with most "similar" tastes.
- With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item)

A key problem of collaborative filtering is how to combine and weight the preferences of user neighbors. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.

Big data has emerged as a widely recognized trend, attracting attentions from government, industry and academia. Generally speaking, Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time" is on the rise. The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. Big Data is still a maturing and evolving discipline. Big data databases and files have scaled beyond the capacities and capabilities of commercial database management systems. Structured representations become a bottleneck to efficient data storage and retrieval. Gartner has noted four major challenges (the four Vs): increasing volume of data, increasing velocity (e.g. in-out and change of data), increasing variety of data types and structures, and increasing variability of data. We have suggested a fifth V: value, which is the contribution big data has to decision making. Add to these the increasing number of disciplines and problem domains where big data is having an impact and one sees an increase in the number of challenges and opportunities for big data to have a major impact on business, science, and government. Concretely, as a critical step in traditional CF algorithms, to compute similarity between every pair of users or services may take too much time, even exceed the processing capability of current RSs. Consequently, service recommendation based on the similar users or similar services would either lose its timeliness or couldn't be done at all. In addition, all services are considered when computing services' rating similarities in traditional CF algorithms while most of them are different to the target service. The ratings of these dissimilar ones may affect the accuracy of predicted rating.

2. Brief description

- *Big data:* Big Data is a broad term for data sets so large or complex that traditional data processing applications are inadequate.
- *Recommender system:* These are a subclass of information filtering system that seek to predict the

’rating’ or ’preference’ that a user would give to an item

- *Clustering:* Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters.
- *Collaborative Filtering:* It is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc.

3. Literature review

A. Community-based user domain model collaborative recommendation algorithm

Collaborative Filtering (CF) is a commonly used technique in recommendation systems. It can promote items of interest to a target user from a large selection of available items. It is divided into two broad classes: memory-based algorithms and model-based algorithms. The latter requires some time to build a model but recommends online items quickly, while the former is time-consuming but does not require pre-building time. Considering the shortcomings of the two types of algorithms, we propose a novel Community-based User domain Collaborative Recommendation Algorithm (CUCRA). The idea comes from the fact that recommendations are usually made by users with similar preferences. The first step is to build a user-user social network based on user’s preference data. The second step is to find communities with similar user preferences using a community detective algorithm. Finally, items are recommended to users by applying collaborative filtering on communities. Because we recommend items to users in communities instead of to an entire social network, the method has perfect online performance. Applying this method to a collaborative tagging system, experimental results show that the recommendation accuracy of CUCRA is relatively good, and the online time-complexity reduces to O (n).

B. A fast clustering algorithm to cluster very large categorical data sets in data mining

Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. However, working only on numeric values limits its use in data mining because data sets in data mining often contain categorical values. In this paper we present an algorithm, called k-modes, to extend the k-means paradigm to categorical domains. We introduce new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency-based method to update modes in the clustering process to minimize the clustering cost function. Tested with the well-known soybean disease data set the algorithm has demonstrated a very good classification performance. Experiments on a very large health insurance data set consisting of half a million records and 34 categorical attributes show that the algorithm is scalable in terms of both the number of clusters and the number of records.

C. Clustering by pattern similarity in large data sets

Clustering is the process of grouping a set of objects into classes of similar objects. Although definitions of similarity vary from one clustering model to another, in most of these models the concept of similarity is based on distances, e.g., Euclidean distance or cosine distance. In other words, similar objects are required to have close values on at least a set of dimensions. In this paper, we explore a more general type of similarity. Under the Cluster model we proposed, two objects are similar if they exhibit a coherent pattern on a subset of dimensions. For instance, in DNA microarray analysis, the expression levels of two genes may rise and fall synchronously in response to a set of environmental stimuli. Although the magnitude of their expression levels may not be close, the patterns they exhibit can be very much alike. Discovery of such clusters of genes is essential in revealing significant connections in gene regulatory networks. E-commerce applications, such as collaborative filtering, can also benefit from the new model, which captures not only the closeness of values of certain leading indicators but also the closeness of (purchasing, browsing, etc.) patterns exhibited by the customers. Our paper introduces an effective algorithm to detect such clusters, and we perform tests on several real and synthetic data sets to show its effectiveness.

4. Clustering

Clustering is a major task in data analysis and data mining applications. It is the method of assigning an objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. The superiority of a clustering result depends on the similarity measure used by the method and its implementation. The superiority of a clustering technique is also calculated by its ability to find out some or all of the hidden patterns. Similarity of a cluster can be expressed by the distance function. In data mining, there are some requirements for clustering the data. Clustering based collaborative filtering approach mainly contains two types of clustering algorithms.

A. Partitioned clustering

Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. The cluster should exhibit two properties, these are (a) each group must contain at least one object (b) each object must belong to exactly one group. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. Partitional clustering contains algorithms like K means clustering, K medoids clustering. But these Partitional algorithms have some limitations.

B. Hierarchical Clustering

Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. A hierarchical method creates a hierarchical decomposition of the given set of data objects. Tree of clusters is called as dendrograms. Every cluster node contains child clusters, sibling clusters partition the points Covered by their common parent.

5. Collaborative filtering

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. Advantage of the collaborative filtering approach is that it does not rely on machine analyzable content. It is capable of accurately recommending complex items without requiring an understanding of the item itself. Collaborative Filtering assumes that people who agree in past will agree in future too and people will like the similar kinds if items they like in the past. Collaborative filtering contains two types of techniques, User based collaborative filtering and Item based collaborative filtering.

A. User based collaborative filtering

User-based collaborative filtering predicts a user's interest in an item which is based on rating information from similar user profiles. User based CF assumes that a good way to find a certain user's interesting item is to find other users who have a similar interest. This type of technique first tries to find the user's neighbors based on user similarities and then combine the neighbor users' rating scores.

B. Item based collaborative filtering

Item based collaborative filtering technique also applies same idea like user based CF but instead of similarity between users it uses similarity between items. The rating of an item by a user can be predicted by averaging the ratings of other similar items rated by user.

6. Conclusion

The computation time taken for processing services gets reduced through clustering. Most similar services can be recommended to user with clustering and collaborative filtering approach.

References

[1] Wanchun Dou, and Jianxun Liu, IEEE Trans. On Club CF, "A Clustering-based Collaborative Filtering Approach for Big Data Application, 2014..

[2] 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.

[3] M. C. Pham, Y. Cao, R. Klamma, et al., "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis," Journal of Universal Computer Science, vol. 17, no. 4, pp.583-604, April 2011.

[4] Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE Big Data, pp. 403-410, October 2013.

[5] J. Mai, Y. Fan, and Y. Shen, "A Neural Networks Based Clustering Collaborative Filtering Algorithm in E-Commerce Recommendation System," in Proc. 2009 Int'l Conf. on Web Information Systems and Mining, pp. 616-619, June 2009.

[6] X. Li, and T. Murata. "Using Multidimensional Clustering Based Collaborative Filtering Approach Improving Recommendation Diversity," in Proc. 2012. IEEE/WIC/ACM Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technology, pp. 169-174, December 2012.

[7] Z. Zhou, M. Sellami, W. Gaaloul, et al., "Data Providing Services Clustering and Management for Facilitating Service Discovery and Replacement," IEEE Trans. on Automation Science and Engineering, vol. 10, no. 4, pp. 1-16, October 2013.