ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 1, pp: (202-208), Month: April 2015 – September 2015, Available at: www.paperpublications.org

Collaborative Filtering Approach for Big Data Applications Based on Clustering

[1]Santhini M, [2]Dr. Balamurugan M, [3]Govindaraj M

[1]M.Tech IT, Bharathidasan University
[2]Associate Professor, Bharathidasan University
[3]Coordinator, Department of Computer Science, Bharathidasan University

Abstract: **Big data deals with large volume of complex growing data set with multiple autonomous sources. With the growing technologies, data storage and data collection capacity goes increases day-by-day, big data are now rapidly expanding in all fields. It tends to increase services on internet. So, the service relevant data become too vast to process by traditional approaches. In the view of these challenges, this survey paper presents HACE theorem, which characterize big data features and Collaborative filtering techniques used in recommender systems. Recommender system is an application deals with information overloaded, used to recommend items to the user. Nowadays Service relevant data become too big to be effectively processed by traditional approaches, so one solution to this challenge is Clustering Based Collaborative Filtering. This approach recruits similar services in the same cluster to recommend services collaboratively.**

Keywords: **Big Data, Collaborative Filtering, Recommender System, HACE.**

I. INTRODUCTION

Data is large volume data. It initiative span for unique dimension: volume, velocity, variety, veracity [3]. Big data concerns large volume complex growing data set with multiple, autonomous sources. Searching on Google for an electronic item, gives number of searches related to that item from various autonomous online sites. This will result in large data generation. As comment and views keep coming on internet from various users for item. Can we summarize all types of opinion and relate it to our choice? All types of opinion in different media in a real time fashion, including updated, cross-referenced discussions by end users. This type of summarization program is an excellent example of big data processing, as information comes from multiple autonomous sources with some of its characteristic. Big data characteristics are useful for discovery of knowledge form big data. They are **H**eterogeneous; **A**utonomous sources with distributed and decentralized control, and **C**omplex and **E**volving relationship among data [2].

1) **H**eterogeneous and Diverse Dimensionality: Big data is heterogeneous, due to different data collector has their own schema or protocols to store information, and nature of different application also results in diverse data representations.

2) **A**utonomous Sources with Distributed and Decentralized Control: It is one of the main characteristic of big data. The autonomous system is able to generate and collect information without involving any centralized control. This is similar to the WWW setting where each web server provides a certain amount of information and each web server able to function without necessarily relying on other server.

3) **C**omplex and **E**volving Relationships: In centralize data storage system, data fields such as age, gender, income, education backgrounds used to represent individual characteristics. These sample features used to treat individual entity independently without considering their social connections. This social connection is one of the most important factors of human society, which includes individual belongings. E.g., our friend circle may form based on the common hobbies or people are connected by biological relationships. Major social network sites, such as Facebook or Twitter characterized by social functions such as friend-connections and followers (in Twitter) [2].

These characteristics of big data and its application induce challenges for data processing and computing, data privacy and algorithms for big data domain. These challenges are explain in brief in following section.

Clustering are such techniques that can reduce the data size by a large factor by grouping similar services together. Therefore, Collaborative Filtering approach based on Clustering , which consists of two stages: clustering and collaborative filtering is proposed. Clustering is a preprocessing step to separate big data into manageable parts. A cluster contains some similar services. As the number of services in a cluster is much less than the total number of services, the computation time of CF algorithm can be reduced significantly. Besides, since the ratings of similar services within a cluster are more relevant than that of dissimilar services, the recommendation accuracy based on users ratings may be enhanced.

II. DATA MINING CHALLENGES WITH BIG DATA TIER

1: Data Accessing and computing this tier focuses on data accessing and arithmetic computing procedures. Because Big Data are often stored in different locations and data volumes may continuously grows. For computing of large distributed data storage, an effective computing platform needed. Data mining algorithms require all data to be loaded into the main memory. However, this is becoming a clear technical barrier for Big Data because moving data across different locations is fine for small data, but if data is vast then it is not possible to load such a large data in main memory. Tier

2: Data privacy and domain knowledge At Tier II, centre on semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). Tier

3: Big Data mining algorithm At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. In the following section, technique for relevant data retrieving and recommending relevant services to the user will be explain. Those are CF (collaborative filtering) and Club-CF (Cluster based collaborative filtering).

III. PROPOSED SYSTEM

Clustering based collaborative filtering approach contains tow modules. First Clustering, in this services are clustered depend on similarity in Description, Functionality & Characteristics respectively. Second Collaborative Filtering, in this, first rating similarity is computed & then predicted rating is given to the clustered services.

a) Clustering:

In step clustering first stem words are recognized by using Porter Stemmer Algorithm. Then similarity in services based on Description & Functionality is Computed By using Jaccard similarity coefficient (JSC). Characteristic similarity between two services is computed by using weighted sum of Description Similarity and Functionality Similarity. At last services are clustered using Agglomerative Hierarchical Clustering Algorithm.

b) Collaborative Filtering:

In this module rating similarity between two services is computed by using Pearson correlation coefficient (PCC).Then neighboring services are selected by using Constraint Formula. In last step all recommended services are ranked in nonascending order according to their predicted ratings.

IV. CLUSTERING

Clustering is a major task in data analysis and data mining applications. It is the method of assigning a 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) *Partitional 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. Hierarchical clustering is further divided in to two types.

Agglomerative: Agglomerative hierarchical clustering is a bottom-up clustering method. It starts by letting each object form its own cluster and iteratively merges cluster into larger and larger clusters, until all the objects are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchies root. For the merging step, it finds the two clusters that are closest to each other, and combines the two to form one cluster. Clustering based collaborative filtering approach uses agglomerative algorithm for clustering services.

Divisive: It works in a similar way to agglomerative clustering but in the opposite direction. As it uses top down approach, this method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain.

V. RECOMMENDER SYSTEM

Recommender systems or recommendation systems are a subclass of information filtering system that used to predict the rating or preference that user would give to an item. Recommender systems typically produce a list of recommendations by using one of two ways - through collaborative filtering or content-based filtering.

VI. COLLABORATIVE FILTERING: PRELIMINARY KNOWLEDGE

The 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. E-commerce and online shopping sites uses Recommender systems. Systems correlate Personal tastes or interests.

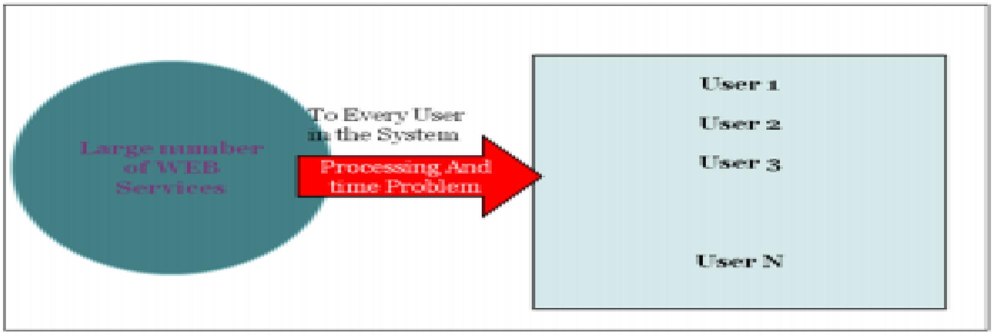
Large number of WEB Services
To Every User in the System
Processing And time Problem
User 1
User 2
User 3
User N

Fig.1 Traditional Approach to recommend services to user

Above figure 1and figure2 depicts the traditional approach and similarity-based approach for recommending services to the user. Similarity-based approach is base of CF based clustering.

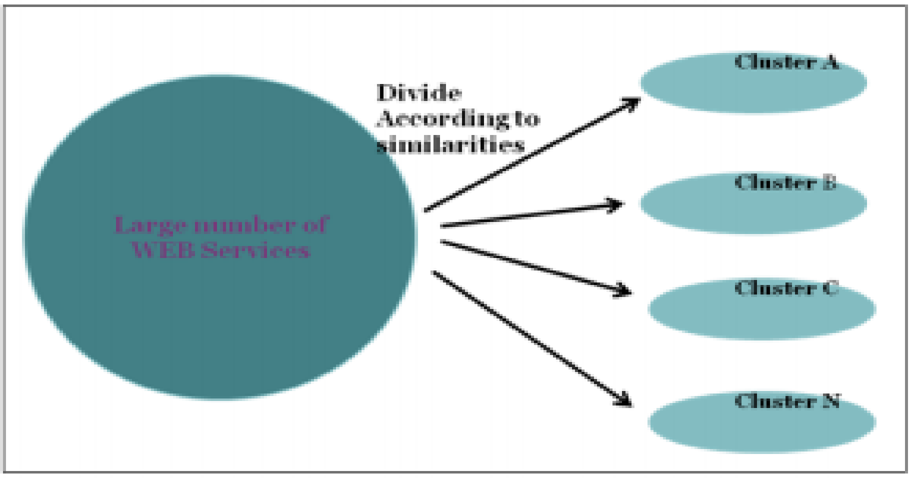
Large number of WEB Services
Divide According to similarities
Cluster A
Cluster B
Cluster C
Cluster N

Fig.2 Similarity based Approach for recommend services to user

a) Collaborative Filtering (CF) And CF Based Clustering:

Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly. For example, a collaborative filtering recommendation 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). Note that these predictions are specific to the user, but use information gleaned from many users. This differs from the simpler approach of giving an average (non-specific) score for each item of interest, for example based on its number of votes. Collaborative filtering contains two types of techniques, User based collaborative filtering and Item based collaborative filtering.

i) 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 neighbour users’ rating scores.

ii) 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.

b) CF Based Clustering:

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. The emerging of service computing and cloud computing, more and more services are deploys in cloud infrastructures to provide rich functionalities.

Service users have nowadays encounter unprecedented difficulties in finding ideal ones from the overwhelming services. Recommender systems (RSs) are techniques and intelligent applications to assist users in a decision making process where they want to choose some items among a potentially overwhelming set of alternative products or services. Collaborative filtering (CF) such as item- and user based methods are the dominant techniques applied in RSs [5]. The basic assumption of user-based CF is that people who agree in the past tend to agree again in the future. Different with user-based CF, the item-based CF algorithm recommends a user the items that are similar to what he/she has preferred before.

Although traditional CF techniques are sound and have been successfully applied in many e-commerce RSs, they encounter two main challenges for big data application: 1) to make decision within acceptable time; and 2) to generate ideal recommendations from so many services. 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. Service recommendation based on the similar users or similar services would either lose its timeliness or couldn't be done at all. When computing services rating similarities in traditional CF algorithms, all services must consider, while most of them are different to the target service. The ratings of these dissimilar ones may affect the accuracy of predicted rating. A naive solution is to decrease the number of services that we need to be process in real time. Clustering are such techniques that can reduce the data size by a large factor by grouping similar services together. Therefore, we propose a Collaborative Filtering Based Clustering, which consists of two stages: Clustering and Collaborative filtering. Clustering is a pre-processing step to separate big data into manageable parts.

VII. LITERATURE SURVEY

Clustering method for CF has studied by some researchers. The cluster analysis gathers users with similar characteristics. This mechanism uses user-rating data to compute similarity between users or items. This used for making service recommendations to user. This was the earlier mechanism used in many commercial systems. There are two basic methods of CF: User–User Collaborative Filtering: User–user CF is a straightforward algorithmic interpretation of collaborative filtering: find other users whose past rating behaviour is similar to that of the current user and use their ratings on other items to predict what the current user will like. To predict Mary's preference for an item she has not rated, user–user CF looks for other users who have high agreement with Mary on the items they have both rated. Item–Item Collaborative Filtering: User–user collaborative filtering, while effective, suffers from scalability problems as the user base grows. To extend collaborative filtering to large user bases and facilitate deployment on ecommerce sites, it was necessary to develop algorithms that are more scalable. Item–item collaborative filtering, also called item-based collaborative filtering, takes a major step in this direction and is one of the most widely deployed collaborative filtering techniques today. Item–item CF uses similarities between the rating patterns of items. If two items belongs to the same users like and dislike then, they are similar and users may have similar preferences for similar items. Following are some CF techniques using clusters.

a) Neural Network-Based CF [4]:

E-commerce recommendation system is one of the most important and the most successful application field of data mining technology. Recommendation algorithm is the core of the recommendation system. In this paper, a neural networks-based clustering collaborative filtering algorithm in e-commerce recommendation system is designed, trying to establish an classifier model based on BP neural network for the pre-classification to items and giving realization of clustering collaborative filtering algorithm and BP neural network algorithm, and carrying on the analysis and discussion to this algorithm from multiple aspects.

- Cluster analysis collects users with similar characteristics according to web visiting message data.
- It may not possible to say that a user's preferences to web visiting are relevant to preference on purchasing.

b) Multi-Dimensional Clustering into CF [5]:

This approach provides a flexible solution that incorporates multidimensional clustering into a collaborative filtering recommendation model to provide a quality recommendation. This facilitates to obtain user clusters, which have diverse preference from multi-view for improving effectiveness and diversity of recommendation. The presented algorithm works in three phases: data pre-processing and multidimensional clustering, choosing the appropriate clusters and recommending for the target user Background data are collected in the form of item and user profiles and clustered using algorithm.

- Clusters are formed on item and user profile. Then, Poor clusters with similar feature were eliminated.
- While appropriate clusters were further selected based on cluster pruning. Item prediction was made.

Disadvantages: This Approach was likely to trade off on increasing the diversity of recommending while maintaining the accuracy.

c) Data Providing Services [6]:

With the increasing number of services available within an enterprise and over the Internet, locating a service online may not be appropriate from the performance perspective, especially in large Internet-based service repositories. Instead, services usually need to be clustered according to their similarity. Thereafter, services in one or several clusters are necessary to be examined online during dynamic service discovery. In this paper, we propose to cluster data providing (DP) services using a refined fuzzy C-means algorithm. We consider the composite relation between DP service elements (i.e., input, output, and semantic relation between them) when representing DP services in terms of vectors. Vectors were clustered using refine fuzzy algorithm. Merging similar services into same cluster, capabilities of services search engines were improved specially in large internet based service repositories. In this approach, it is assumes that domain ontology exist for facilitating semantic interoperability. Disadvantages: Not suitable were lake of parameter exists. 4.4

d) Network Clustering Technique on Social Network [7]:

Collaborative Filtering (CF) is a well-known technique in recommender systems exploits relationships between users and recommends items to the active user according to the ratings of his/her neighbours. CF suffers from the data sparsity problem, where users only rate a small set of items. That makes the computation of similarity between users imprecise and consequently reduces the accuracy of CF algorithms. In this article, we propose a clustering approach based on the social information of users to derive the recommendations. We study the application of this approach in two application scenarios: academic venue recommendation based on collaboration information and trust-based recommendation.

- To identify users' neighbourhood.
- Then use the traditional CF algorithms to generate there commendations.
- This work depends on social relationships between users.

VIII. EVALUATION OF LITERATURE SURVEY

Table.1 Evolution of Literature Work

Parameters **Methods**	**Clustering Basis**	**Accuracy**	**Suitable for providing RS**	**Requirements**
NN-CF	Web Visiting Message Data	Poor	Not Always	Active User Participation
MD- Clustering	User and Item Profile	Poor	Not Always	Active User Participation
DP-Service	Vectors	Poor	Not Always	Active User Participation
N/W clustering on Social N/W	Social Relationships between User	Improved	Yes	Active User Participation

IX. CONCLUSION

Collaborative Filtering Based Clustering approach is relevant to service recommendation. Before applying CF technique, services are merged into some clusters via an AHC algorithm. Then the rating similarities between services within the same cluster are computed. As the number of services in a cluster is much less than that of in the whole system, CF based Clustering costs less online computation time. First, in the respect of service similarity, semantic analysis may be done on the description text of service. In this way, similar services can be clustered together, which will increase the coverage of recommendations. Second, with respect to users, mining their implicit interests from usage records or reviews may be a complement to the explicit interests (ratings). By this means, recommendations can be generated even if there are only few ratings. Social media can be involved in such approach. By many factor similarities based approach for big data applications is relevant to service recommendation. Before applying CF technique, services are merged into some clusters. Then the rating similarities between services within the same cluster are computed. As the number of services in a cluster is much less than that of in the whole system, Club CF costs less online computation time.

X. FUTURE WORK

Future research can be done in two areas. First, in the respect of service similarity, semantic analysis may be performed on the description text of service. In this way, more semantic-similar services may be clustered together, which will increase the coverage of recommendations. Second, with respect to users, mining their implicit interests from usage records or reviews may be a complement to the explicit interests (ratings). By this means, recommendations can be generated even if there are only few ratings. This will solve the sparsely problem to some extent.

REFERENCES

[1] Wanchun Dou*, Member, IEEE, Jianxun Liu, Member, IEEE Trans. On ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application Rong Hu, Member, IEEE, 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] 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.

[4] J. Mai, Y. Fan, and Y. Shen, “A Neural NetworksBased 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.

[5] 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.

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

[7] 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.

[8] G. Thilagavathi, D. Srivaishnavi, N. Aparna, et al., “A Survey on Efficient Hierarchical Algorithm used in Clustering,” International Journal of Engineering, vol. 2, no. 9, September 2013.

[9] J. Wu, L. Chen, Y. Feng, et al., “Predicting quality of service for selection by neighbourhood-based collaborative filtering,” IEEE Trans. on Systems, Man, and Cybernetics: Systems, vol. 43, no. 2, pp. 428-439, March 2013.

[10] 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.

[11] R. D. Simon, X. Tengke, and W. Shengrui, “Combining collaborative filtering and clustering for implicit recommender system,” in Proc 2013 IEEE 27th Int’l Conf. on. Advanced Information Networking and Applications, pp. 748-755, March 2013.