ISSN:0975-9646

Collaborative Filtering Approach For Big Data Applications in Social Networks

Rajeswari.M

Assistant Professor, Department of Information Technology,Velammal Institute of Technology,Anna University

Chennai, Tamilnadu, India

Abstract: In this paper Collaborative Filtering approach (Club CF) is planned recruiting similar services in the same clusters to recommend services collaboratively (Recommender Service System).Technically this approach is enacted around two stages: In first stage, the available services are divided into small-scale clusters in logic for further processing. In second stage, a collaborative filtering algorithm is imposed on one of the clusters. The number of the services in a cluster is much less than the total number of the services available on the web, it is used to reduce the online execution time of collaborative filtering.

Keywords: Big data application, cluster, mash up techniques, collaborative filtering approach

I. INTRODUCTION

Big data has been widely recognized trend, that attracting attentions from government, academia an. Generally speaking, Big Data concerns and industry, large-volume, complex, growing data sets with multiple[1][2], autonomous sources. Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture the data 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 no off data and extract useful information or knowledge for future actions. With the prevalence of service computing and cloud computing, more and more services are deployed in cloud infrastructures to provide rich functionalities. Service users have nowadays encounter unprecedented difficulties in finding ideal ones from the overwhelming services. Recommender systems [6] are techniques and intelligent applications to assist users in a decision making process where they want to and past purchases. Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.Recalculate entire risk portfolios in minutes. Quickly identify customers who matter the most things are click stream analysis and data mining to detect fraudulent behaviour. Collaborative filtering (CF) such as item- and user-based methods are the dominant techniques applied in RSs. 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 commerce RSs, they encounter two main challenges for big data application:

- Make decision within acceptable time
- Generate ideal recommendations from so many services.

timeliness or could

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.

II.BACKGROUND

Data mining an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Traditional data mining algorithms when applied on these Big data results in poor performance with respect to the computational part. So, there is a need to parallelize the traditional data mining algorithms. There has been several research works carried out to handle and process the Big data. Google has developed a software framework called Map Reduce to support large distributed data sets on clusters of computers, which is effective to analyse large amounts of data. Followed by Google’s work, many implementations of Map Reduce emerged and lots of traditional methods combined with Map Reduce have been presented such as Apache Hadoop, Phoenix, Mars, and Twister. Apache Hadoop is a software framework that helps constructing the reliable, scalable distributed systems. Hadoop enables users to store and process large volumes of data and analyse it in ways not previously possible with less scalable solutions or standard SQL-based approaches. In our work, we have discussed the various advantages of incorporating parallelism in existing mining algorithms and proposed a system for mining Big data.

Service recommendation [6] based on the similar users or similar services would either lose its not 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.

A. Methods

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources so existing knowledge particular area only handled. 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 may Possible[5].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 could not be done at all because decided also business people only. Existing neural networks-based clustering algorithm in e-commerce recommendation system. The cluster analysis gathers users with similar characteristics according to the web visiting message data. it is hard to say that a users preference on web visiting is relevant to preference on purchasing. The vectors were clustered using a refined fuzzy C-means algorithm. Through merging similar services into a same cluster, the capability of service search engine was improved significantly, especially in large Internet-based service repositories. this approach is not suitable for some services which are lack of parameters.

B. Limitations

✓ *K*-means clustering algorithm was applied to partition movies based on the genre requested by the user.

✓ The clusters with similar features were deleted while the appropriate clusters were further selected based on cluster pruning.

✓ Much Time to Search and data Clustering Management is poor performance for the relevant data publish.

✓ Top Ranking Sites Only Published From the some Mining Algorithms Basis

C.Techniques

Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture the image and process within a “tolerable elapsed time” is give the Solution. Our application Knowledge of service computing and cloud computing, more and more services are deployed in cloud infrastructures to provide rich functionalities [7].The description and functionality information is considered as metadata to measure the characteristic similarities between services. According to such similarities, all services are merged into smaller-size clusters. Then CF algorithm is applied on the services within the same cluster. Compared with the above approaches, this approach does not require extra inputs of users and suits different types of services. The clustering algorithm used in ClubCF need not consider the dependence of nodes. 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. 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.

✓ Recommender Service System (RSS)

✓ Massive Parallel Processing (MPP)

✓ Map Reduce

✓ Hadoop

✓ Collaborative Filtering approaches

D.Advantages

✓ This application no need SQL Queries to Users need for relevant searches in Web develop application.

✓ Mashup services from Programmable Web development application is adopt the Big data Cluster Filtering Knowledge.

✓ Recommender Service System is Provide our knowledge provide Based from the users preferences.

✓ Avoid Reduplication Data’s and improve performance the cluster Filtering

1. Recommender Service System

we present a Recommended approach for big data applications 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, ClubCF costs less online computation time. Moreover, as the ratings of services in the same cluster are more relevant with each other than with the ones in other clusters, prediction based on the ratings of the services in the same cluster will be more accurate than based on the ratings of all similar or dissimilar services in all clusters. These two advantageous been verified by experiments on real-world data set.

2. *Collaborative Filtering*

ClubCF spends less computation time than Item-based CF. Since the number of services in a cluster is fewer than the total number of services, the time of rating similarity computation between every pair of services will be greatly reduced. As the rating similarity threshold γ increase, the computation time of ClubCF decrease. It is due to the number of neighbors of the target service decreases when increase. However, only when $\gamma = 0.4$, the decrease of computation time of IbCF is visible. It is due to the number of neighbors found from a cluster may less than that of found from all, and then it may spend less time on computing predicted ratings in Club CF. as increase, the computation time of Club CF decrease obviously.

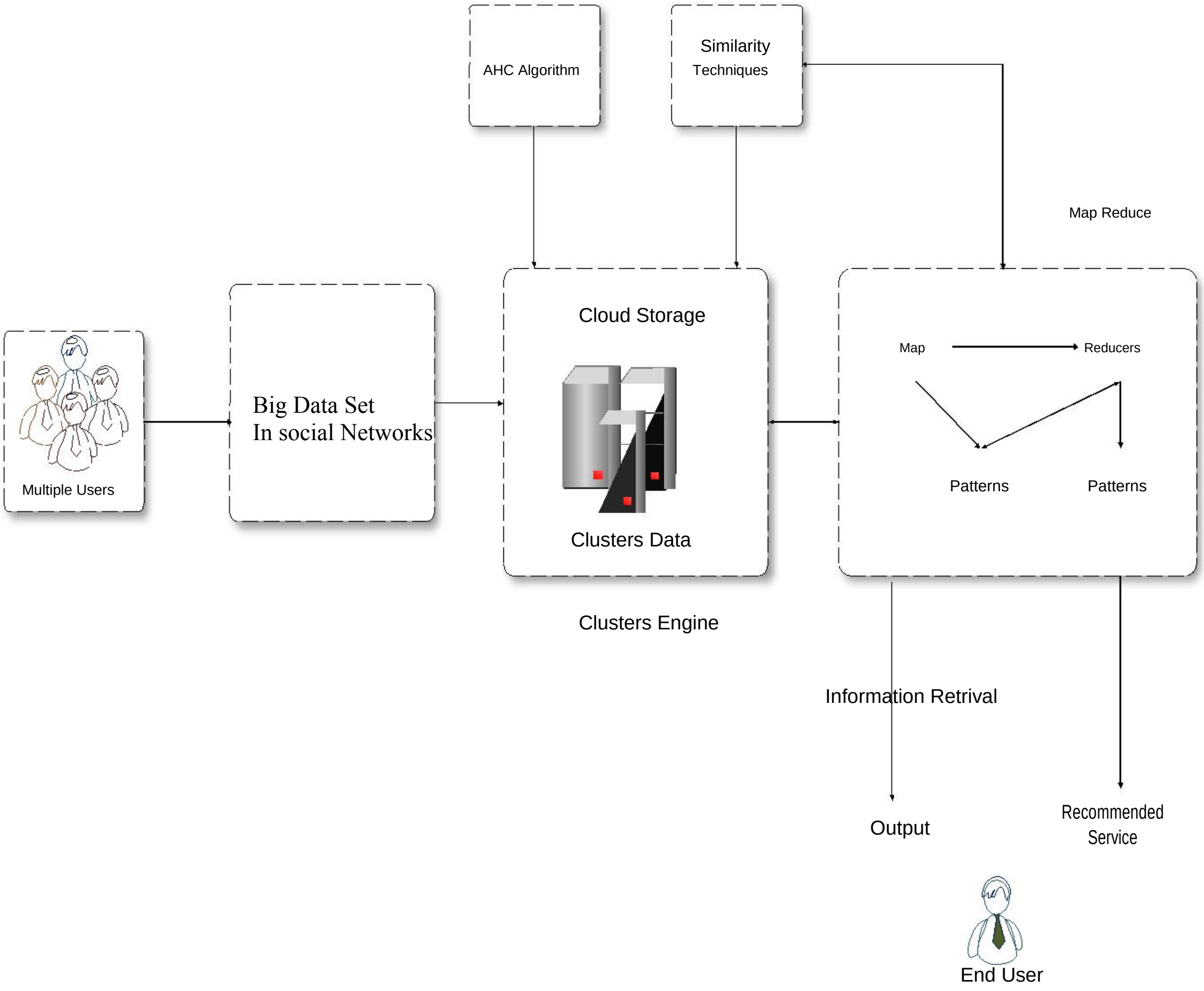

Fig. 1 Architecture of collaborative filtering algorithm

Since a bigger *K* means fewer services in each cluster a bigger γ makes less neighbours, the computation time predicted ratings based on less neighbours may decrease.ClubCF is a revised version of tradition alitem-based CF approach for adapting to big data environment. Therefore, to verify its accuracy, we compare the MAE ClubCF with a traditional item-based (IbCF) described each test mash up service in each fold, its predicted is calculated based on IbCF and ClubCF approach separately. To evaluate the accuracy of ClubCF, into Absolute Error (MAE), which is a measure of the deviation of recommendations from their true user-specified ratings?

3. *Mash up Techniques*

Mash up is a term that's become popular to describe Web 2.0-ish sites that combine the features or functions of one website with another [8]. But the term mash up has its roots in music, where creative (or bored) people combine the vocal and instrumental tracks from two or more songs to create a new song. Sometimes the point is simply humour, created by the contrast of different musical genres. This Wikipedia entry gives more of the history, and many examples of musical mash ups.But the focus of this article Service, and is on website mash ups, created by clever programmers of who use a variety of techniques to create useful new services that are derived from existing ones. These sites typically feature a high level of interactivity, user input, social networking, and sometimes even encourage people of to use them as the basis for derivative works.

4. *Mapping Mash Ups*

When the very popular Google Maps released an API that Mean allowed web developers to easily integrate mapping

their own sites, it spurred a lot of creative minds into action. APIs for Yahoo Maps, MapQuest and Microsoft's Virtual Earth shortly followed, making it almost trivial to plug a rich source of geographical, topological, street-level and satellite image data into existing websites. Here are some examples of interesting map mash ups Ask 500 People is an innovative survey tool designed to gather opinion data in minutes instead of days. Business and individuals can pose a question, and a diverse, decentralized pool of people from around the world will be polled to provide their answer or opinion. The voting results are dynamically displayed on a world map that's powered by Google Maps. Weather Bonk combines Google Maps with data from the National Weather

Weather Underground, and a collection of personal weather stations in homes and schools. You can also view live webcam images from the area you choose.

5. *Video And Photo Mash Ups*

Flickr, Yahoo Photos and You tube. Combine those with keyword tagging data There are an abundance of video and photo hosting sites with APIs, such as that not only specifies the subject matter of the media, but also the geographic location of the imagery, and you can come up with some pretty cool mash ups. You specify a set of keywords, and Magnify will find relevant videos from YouTube, Google Video, Yahoo, Revver, Blip and other sources. You can also add community features to your site, such as peer review, comments and discussion forums. Flickr Sudoku lets you play Sudoku online, using images of numbers from Flickr.Flickr Maps is a mash up of Yahoo Maps and Flickr that turns you into a virtual tourist. Select a city in the USA, and then view relevant photos from Flickr.

6. *News Mash Ups*

RSS feeds are used by many blogs and news organizations as a means of distributing (or syndicating) news headlines and story summaries. So that makes it possible for mash up developers to create personalized newspapers that meet your interests. Other mash ups combine keyword from news stories with maps and photo sharing sites. Here are some notable news mash ups AP News + Google Maps displays news stories from the Associated Press (National, Sports, Business, Technology and Strange) superimposed on Google Maps. The geographic location for each news story is determined using Yahoo's Decoding API, and is then plotted on the map Flicker Fling lets you select a news source, such as CNN or Wired News, and view the latest news rendered in pictures.

7. *Massive Parallel Processing*

Massive analysis algorithms have been utilized where the huge data are stored. Clustering algorithms can be either hierarchical or partitioned. Some standard partitional approaches (e.g., *K*-means) suffer from several limitations:

✓ results depend strongly on the choice of number of clusters *K*, and the correct value of *K* is initially unknown;

✓ cluster size is not monitored during execution of the *K*-means algorithm, some clusters may become empty ("collapse"), and this will cause premature termination of the algorithm;

✓ Algorithms converge to a local minimum.

Hierarchical clustering methods can be further classified into agglomerative or divisive, depending on whether the clustering hierarchy is formed in a bottom-up or top-down fashion.Mancurrentstate-of-the-art clustering systems exploit agglomerative hierarchical clustering (AHC) as their clustering strategy, due to its simple processing structure and acceptable level of performance. Furthermore, it does not require the number of clusters as input. Therefore, we use an AHC algorithm for service clustering as follow.

A. Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

These bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces [10]. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.

The technique is also used to measure cohesion within clusters in the field of data mining.

Cosine distance is a term often used for the complement in positive space, that is:

$$D_C(A,B) = 1 - S_C(A,B).$$

It is important to note, however, that this is not a proper distance metric as it does not have the triangle inequality property and it violates the coincidence axiom; to repair the triangle inequality property whilst maintaining the same ordering, it is necessary to convert to Angular distance.

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

Given two vectors of attributes, *A* and *B*, the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

✓ For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

✓

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf- idf weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

✓

If the attribute vectors are normalized by subtracting the vector means (e.g., $A - \bar{A}$), the measure is called centred cosine similarity and is equivalent to the Pearson Correlation Coefficient.

III.Conclusion

In attendance a Club CF approach for big data applications relevant to service suggestion. Proceeding to 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, Club CF costs less online computation time. Moreover, as the ratings of services in the same cluster are more relevant with each other than with the ones in other clusters, prediction based on the ratings of the services in the same cluster will be more accurate than based on the ratings of all similar or dissimilar services in all clusters.

IV. 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] M. A. Beyer and D. Laney, "The importance of "big data": A definition," Gartner, Tech. Rep., 2012.

[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] A. Rajaraman and J. D. Ullman, "Mining of massive datasets,"
Cambridge University Press, 2012.

[4] 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.

[5] A. Bell, I. Cantando, F. Diaz, et al., "An empirical comparison of social, collaborative filtering, and hybrid recommenders," *ACM Trans. On Intelligent Systems and Technology*, vol. 4, no. 1, pp. 1-37, January 2013.

[6] W. Zing, M. S. Shang, Q. M. Zhang, et al., "Can Dissimilar Users
Contribute to Accuracy and Diversity of Personalized Recommendation?," *International Journal of Modern Physics C*, vol. 21, no. 10, pp.1217-1227, June 2010.

[7] T. C. Havens, J. C. Bedeck, C. Lecky, L. O. Hall, and M. Palana swami, "Fuzzy c-Means Algorithms for Very Large Data," *IEEE*
Trans. on Fuzzy Systems, vol. 20, no. 6, pp. 1130-1146, December 2012.

[8] Z. Liu, P. Li, Y. Zhen, et al., "Clustering to find exemplar terms for key phrase extraction," in *Proc. 2009 Conf. on Empirical Methods in Natural Language Processing*, pp. 257-266, May 2009.

[9] X. Liu, G. Huang, and H. Mei, "Discovering homogeneous web service community in the user-centric web environment,"*IEEE Trans. On Services Computing*, vol. 2, no. 2, pp. 167-181, April-June 2009.

[10] Z. Liu, P. Li, Y. Zhen, et al., "Clustering to find exemplar terms for key phrase extraction," in Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 257-266, May 2009.

[11] M. J. Li, M. K. Ng, Y. M. Cheung, et al. "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters,"
IEEE Trans. on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1519-1534, November 2008.