International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Clustering Based Hybrid Recommendation System for Services in Big Data

Powar Gayatri Ashok[1], Dr. D. M. Yadav[2]

[1] Department of Computer Engineering, JSPM's Rajarshi Shahu School of Engineering & Research, Narhe, Pune, India

[2] Director, JSPM's Rajarshi Shahu School of Engineering & Research, Narhe, Pune, India

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. It becomes difficult for users to select best product from so many products which are available. These systems suffer from scalability, data sparsity, and cold-start problems resulting in poor quality recommendations. In order to view this problem this paper provides a hybrid recommendation system which will satisfy the users according to their needs and interest and increase the overall performance of the system. The main idea is using hybrid recommendation techniques to suppress the drawbacks of the traditional techniques or an individual technique in a combined model. Paper presents a system to improve the accuracy of recommendation in big data application.***

Keywords: Big data, Clustering, Collaborative Filtering, Hybrid Recommendation system.

1. Introduction

Data is large volume data. There are some characteristics of big data as: volume, velocity, variety, veracity[1] [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. Big data characteristics are useful for discovery of knowledge form big data. They are heterogeneous; **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.

As the development of the internet, intranet and electronic commerce systems, there are amounts of information arrived we can hardly deal with. So, other recommendation systems are available to give useful data or information filtering technologies. Information filtering consist of two main techniques. One is the content based filtering and the other is the collaborative filtering. In terms of theory and implementation Collaborative filtering (CF) has proved one of the most effective technique[1][2]. Different researchers may use different kinds of CF technologies to make recommendation which will have quality. All of them make a use of same data structure in order to provide recommendation as user-item matrix having users and items consisting of their rating scores. It consist of two methods in CF as user based collaborative filtering and item based collaborative filtering [3,4].

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. So, at first, it tries to find the user's neighbors based on user similarities and then combine the neighbor users rating scores, which have previously been expressed, by similarity weighted averaging. Item based collaborative filtering and the user based collaborative filtering uses the same technique. It looks into a set of items; the target user has given rating already and computes how similar they are to the target item under recommendation. After that, it also combines his previous preferences based on these item similarities.

Limitaions of Collaborative Filtering are as following :[5]

Cold-start:
When a new user or item is going to enter into the system there is very few information available about that newly added user or item so based on such limited information it becomes difficult to provide recommendations. This problem applies to new items and is particularly detrimental to users with eclectic interest. Likewise, a new user has to rate a sufficient number of items before the CF algorithm be able to provide accurate recommendations.

Sparsity:
Even as users are very active, there are a few rating of the total number of items available in a user item ratings database. As the basic of the collaborative filtering algorithms are based on similarity measures computed over the co-rated set of items, large levels of sparsity can lead to less accuracy.

2. Related Work

ClubCF Approach [6]
Paper proposes Clustering based Collaborative Filtering i.e ClubCF approach which works in the two stages clustering and collaborative filtering stage. Reduces execution time. Clustering methods are used to divide a set of objects into clusters such that objects in the same cluster are more similar with one another than objects in different clusters. Item based collaborative filtering is imposed on those clusters.

An Efficient Hybrid Algorithm for Clustering [7]
Clustering techniques have received attention in many fields The K-means algorithm is one of the most common techniques used for clustering. However, the results of K-means depend on the initial state and converge to local optima. In order to overcome local optima obstacles, This paper presents an efficient hybrid evolutionary optimization algorithm based on combining Modify Imperialist Competitive Algorithm (MICA) and K-means (K), which is called K-MICA for optimum clustering N objects into K clusters

Bigtable [8]
Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. The paper describe the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and describe the design and implementation of Bigtable.

Service Generated Big Data & Big DaaS [9]
More and more services are emerging on the Internet. Such kind of service-generated data become too large and complex to be effectively processed by traditional approaches. How to store, manage, and create values from the service-oriented big data become an important research problem. To address this challenge, this paper provides an overview of service-generated big data and Big Data-as-a-Service.

Prediction of QoS values of web services [10]
With increasing presence and adoption of Web services on the World Wide Web, Quality-of-Service (QoS) is becoming important for describing nonfunctional characteristics of Web services. In this paper, we present a collaborative filtering approach for predicting QoS values of Web services and making Web service recommendation by taking advantages of past usage experiences of service users.

Neural Network-Based Club-CF [11]
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.

Data Providing Services [12]
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. This paper proposes a cluster data providing (DP) services using a refined fuzzy C-means algorithm.

Network Clustering Technique on Social Network [13]
Collaborative Filtering (CF) is a well-known technique in recommender systems. 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. This article proposes a clustering approach based on the social information of users to derive the recommendations.

3. Existing System

The Existing System approach is divided in two stages [11].

3.1 Clustering Stage

The total number of services is divided into small-scale clusters. Clustering is a critical step. Clustering methods are used to divide a set of objects into clusters such that objects in the same cluster are more similar with one another than objects in different clusters. Cluster analysis algorithms are used where the large data are stored [14][15]. Clustering algorithms can be either hierarchical or partitional. Many current state-of-the-art clustering systems uses the Agglomerative Hierarchical Clustering (AHC)[16] as their clustering strategy, because structure processing technique is simple and have good acceptable level of performance. Furthermore, it does not require the number of clusters as input. Here AHC algorithm is used for service clustering.

3.2 Collaborative Filtering stage

A collaborative filtering algorithm is imposed on one of the clusters. Here the item based collaborative filtering is applied on the clusters in order to provide recommendations.

Problem Definition:

- Recommender System is used to a collection of users for items or products in order to generate meaningful recommendations according to their requirements.
- Improve the speed of recommendation in big data application and also reduce the human efforts of doing analysis process while searching products online by providing recommendations.

4. Proposed System

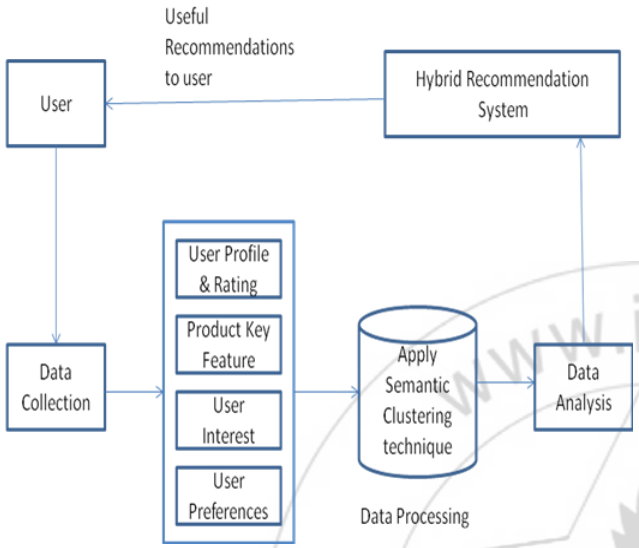

Figure 1: Proposed Architectural Design

The main idea of the proposed system is to use the Hybrid Recommendation System to provide useful recommendations to a collection of users or customers for items or products that might interest them or according to customer's preferences or area of interest; the Hybrid recommendation system combines two or more recommendation techniques in order to increase the overall performance. Here the hybrid recommendation system is going to combine the following two techniques: Content based technique and knowledge based technique. The main idea is using multiple recommendation techniques to suppress the drawbacks of an individual technique in a combined model.

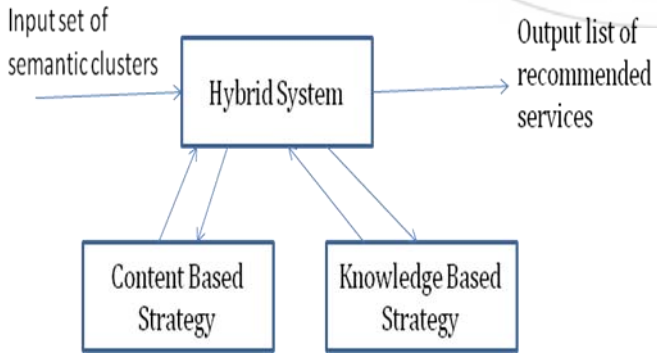

Figure 2: Hybrid System.

In the above architectural design diagram recommendations are provided to user by using clustering based hybrid recommendation system for semantic clusters. First of all the user interact with data processing unit. Data is collected from users in the form of big data; data will be in the form of users purchase information, rating information, according to users area of interest, product key features, purchase history. After the data collection data processing step takes place where semantic clustering techniques are applied to collected data. Complete data is divided into the number of clusters. For this purpose Agglomerative Hierarchical Cluster (AHC) algorithm is used[14]. These clusters are passed to the collaborative filtering techniques. Hybrid recommendation system is used to provide final recommendations to the users.

4.1 Content Based Recommendation Algorithm

A content based recommendation Algorithm works on the user preferences that is likes and dislikes given by any user to items or products and the user profile. Here it will only consider likes given by user .The output of this algorithm will be a set of products. The algorithm is as shown below.

Input: Item wise users preference (Likes/ Dislikes)
List< Up >; User profile< Uf >

Output: Return minimal conflict set of products

Procedure:
Function FCBPred(List< Up > objPre , List< Uf > objProfiles)
Let,
List< temProd > = ∅ // Temporary list for product.
If (objProd ==∅ or objPre == ∅ or objProfiles == ∅) then // Check if Input data is null ?
return ∅
End If
For each Uf In objProfiles
Then // Travel user Profiles
For each Pobj In objProd Then // Travel Product key feature list
If List< Nu > =FindNearestNeighbourUser(Uf , List< Uf >)
Then // Identify the user neighbors i.e same in category
List<itemProd> = Add(CollectTheItemByPreferances(List< Nu >, List< Up >)) // Compute the list of items by preferences given by neighbor user.
End If;
Next;
Return List< temProd >;
End Function;

4.2. Knowledge Based Recommendation Algorithm:

A knowledge based recommendation algorithm works on the set of requirements of user and the product description. Products feature and category are compared with the users interest and category respectively. Output will be a set of products. The following algorithm represents a knowledge based recommendation algorithm.

Input: trusted knowledge (items) List< Pd >; Set of requirements List< Uin >

Output: Return minimal conflict set of products

Procedure:

Function FKBPred(List< Pd > objProd , List<Uin> objInt)
Let,
List< temProd > = ∅ // Temporary list for product.
If (objProd == ∅ or objInt == ∅) then // Check if Input data is null ?
return ∅
End If
For each Uin In objInt Then // Travel user interest list
For eachPobj In objProd Then // Travel Product key feature list
If (Pobj−> KeyFeature == Uin−> Requirement or Pobj−> Category == Uin−> Category)
Then //Compare user interest with product's key feature, categorize
List< temProd > –>Add(Pobj)
EndIf;
End If;
Return List< temProd >;
End Function.

4.3. Hybrid Algorithm

Hybrid algorithm is a combination of content based recommendation algorithm and knowledge based algorithm. The following algorithm represents a hybrid algorithm.

Start Algorithm
Let S = A user session $\{S_1, S_2, \ldots, S_n\}$ Where each S_i has its own Pair of < Product , Attribute, Interest, Hobbies, View's, Product Purchase history, Product likes etc>

List<Rec_Products> R_{obj} - Return list of recommended product's to respective users session S_i.
P_d - Product with its key features or description
U_f - User profile
P_r - Product rating's given by individual users
U_{in} - User's attribute's and area of interest.
U_p -User Preferences' (Likes / Dislikes)
S_{clust} - List of semantic cluster's

For Each SCi In List<S_{clust}>
R_{obj} -> Add (F_CBPred (SC_i -> U_f, SC_i -> U_p)) ; - A content based recommendation using user profile U_f, and user preferances (likes/dislike)U_p.
R_{obj} -> Add (F_KBPred (SC_i -> P_d , SC_i -> U_{in})); - A Knowledge based recommendation on users interest or requirement U_{in}, with product key features P_d .
Next;
Return $R_{obj;}$
Stop Algorithm.

4.4 Mathematical Module

Let HRS be a Hybrid Recommendation System
HRS={ I, S, S_{clust} , R_{obj} , F_CBPred, F_KBPred, P_d, U_f, P_r, U_{in}, U_p, O}
Where,I=Input – big data-Service Clusters
S=User session={S_1, S_2,.... ,S_n}
S_{clust} - List of semantic cluster's
List< S_i >= AHC (Big_{data})
F_CBPred= Content based recommendations on U_f and U_p
F_CBPred=(S_{clust} -< U_f >, S_{clust} - < U_p >)
F_KBPred= Knowledge based recommendations on U_f , U_{in} and P_d.
F_KBPred=(S_{clust} -< U_f >, S_{clust} -< U_{in} >, S_{clust} -< P_d >)
P_d - Product with its key features or description
U_f - User profile
P_r - Product rating's given by individual users
U_{in} - User's attribute's and area of interest.
U_p - User preferences
R_{obj} - Recommended product
O= output - Recommended list R.
R_1 =F_CBPred , R_2 =F_KBPred
R =Σ (R_1 + R_2)
Return R.

5. Experimental Setup and Results

To carry out the experiment we have installed Windows 7 professional 64 Bit and IIS7.0 (Internet Information Services) web server. And both Front-End (Visitor) and Back-End (Admin) site's are published on IIS7.0 by creating two different virtual directories. Using this setup we can able to access and test both site's on Intranet.

$$MAE = \frac{\sum_{i=1}^{n} |r_{a,t} - P(u_a, s_t)|}{n}$$

In order to calculate accuracy Mean Absolute Error (MAE) is calculated as shown in the above equation. Where n is the total no. of items or products. $r_{a,t}$ is the rating given by user Ua to the product. $Pr(Ua - St)$ is the predicted ratings of Ua and St. $Pp(Ua - St)$ is the predicted preferences of Ua and St. $PI(Ua - St)$ is the predicted interested category of Ua and St. Low MAE values represent high accuracy. Calculated MAE values are represented as below. Graphical representation is as shown in the following figure.

$$P_{u_a,s_t} = \overline{r}_{s_t} + \frac{\sum_{s_j \in N(s_t)} (r_{u_a,s_j} - \overline{r}_{s_j}) \times R\_sim'(s_t, s_j)}{\sum_{s_j \in N(s_t)} R\_sim'(s_t, s_j)}$$

Table 1: Comparison between existing system and proposed system

Cluster Size	MAE(Item-Based Recommendation)	MAE(Hybrid Recommendation)
K1(36)	0.666	0.138
K2(80)	0.787	0.225
K3(144)	0.972	0.380
K4(288)	0.958	0.490

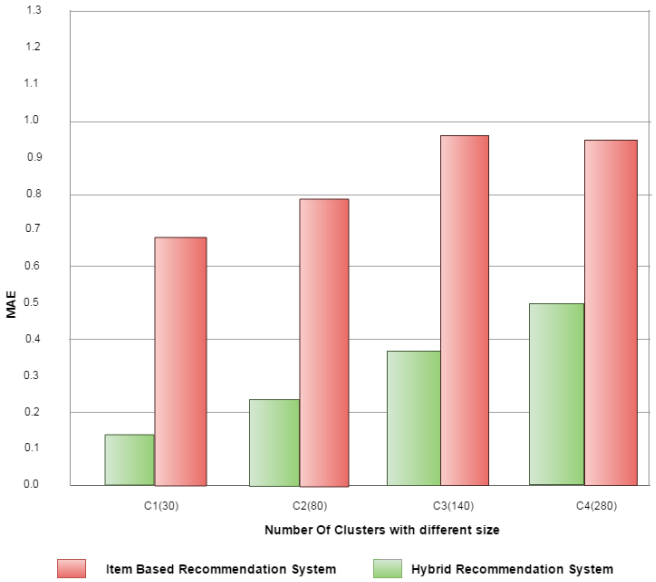

Figure 3: Graphical Representation of Accuracy

6. Conclusion and Future

The Hybrid Recommendation System approach for big data applications is proposed to generate meaningful recommendations to a collection of users for items or products that might interest them. 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, it costs less online computation time. Proposed approach overcomes the limitations of existing system like data sparsity, scalability, accuracy, cold-start problem. Provides recommendations to the customers by improving the accuracy of recommendation in big data application.

For future research another collaborative filtering techniques can be combined to provide even more accurate results. Semantic analysis may be done on the description text of service that is with respect to service similarity.

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, andW. Ding, ``Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97_107, Jan. 2014.
[3] A. Rajaraman and J. D. Ullman, Mining of Massive Datasets.Cambridge, U.K.: Cambridge Univ. Press, 2012.
[4] M. A. Beyer and D. Laney, "The importance of big data: A definition," Gartner, Tech. Rep., 2012.
[5] Comparative Analysis of Collaborative Filtering Technique Urmila Shinde1, Rajashree Shedge, (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 10, Issue 1
[6] 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.
[7] T. Niknam, E. Taherian Fard, N. Pourjafarian, and A. Rousta, ``An efficient algorithm based on modi_ed imperialist competitive algorithm and K-means for data clustering," Eng. Appl. Artif. Intell., vol. 24, no. 2, pp. 306_317, Mar. 2011.
[8] F. Chang et al.,``Bigtable: A distributed storage system for structured data," ACM Trans. Comput. Syst., vol. 26, no. 2, pp. 139, Jun. 2008.
[9] 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.
[10] Z. Zheng, H. Ma, M. R. Lyu, and I. King, ``QoS-aware web service recommendation by collaborative ltering," IEEE Trans. Services Comput.,vol. 4, no. 2, pp. 140152, Feb. 2011.
[11] 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.
[12] 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.
[13] 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.
[14] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE , A. Alamri, I.Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING-2014.
[15] Haytham et al., "Automatic Clustering of e-Commerce Product Description", Journal of Applied Computer Science & Mathematics-2012
[16] R. S. Sandeep, C. Vinay, and S. M. Hemant, ``Strength and accuracy analysis of af_x removal stemming algorithms," Int. J. Comput. Sci. Inf. Technol., vol. 4, no. 2, pp. 265_269, Apr. 2013.