A Clustering Based Hybrid Recommendation System for Services in Big Data

Powar Gayatri Ashok¹, Dr. D. M. Yadav²

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; Autonomous sources with distributed and decentralized control, and Complex and Evolving relationship among data [2].

1) Heterogeneous 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) Autonomous 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) Complex and Evolving 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.
2. Related Work

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 describes the simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and describes 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. Therefore, 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.

3. Existing System

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. Item based collaborative filtering is imposed on those clusters.
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](image)

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.](image)

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 // 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:
is the predicted ratings of . Low MAE values represent high accuracy.

\[ F_{CBPred} = \text{Content based recommendations on } U \]

\[ F_{KBPred} = \text{Knowledge based recommendations on } U_f, U_i \]

For each Pobj In objProd Then // Travel Product key feature

End If;

Return List< temProd >;
End Function.

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.

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. \( \tau_{2,t} \) is the rating given by user \( U_a \) to the product. \( P_r(U_a - St) \) is the predicted ratings of \( U_a \) and \( St \). \( P_p(U_a - St) \) is the predicted preferences of \( U_a \) and \( St \).\( P_l(Ua - St) \) is the predicted interested category of \( U_a \) and \( St \). Low MAE values represent high accuracy. Calculated MAE values are represented as below. Graphical representation is as shown in the following figure.

\[
P_{d,r} = \bar{T} + \sum_{i \in N(y)} \left( \frac{U_{a,d} - \tau_{x,y}}{R_{sim}^2(s_t, s_j)} \right) \sum_{j \in N(y)} R_{sim}^2(s_t, s_j)
\]

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>MAE(Item-Based Recommendation)</th>
<th>MAE(Hybrid Recommendation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1(36)</td>
<td>0.666</td>
<td>0.138</td>
</tr>
<tr>
<td>K2(80)</td>
<td>0.787</td>
<td>0.225</td>
</tr>
<tr>
<td>K3(144)</td>
<td>0.972</td>
<td>0.380</td>
</tr>
<tr>
<td>K4(288)</td>
<td>0.958</td>
<td>0.490</td>
</tr>
</tbody>
</table>
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


