A Collaborative Filtering Approach for Service Based Recommender Systems using Big Data

1P. Deepika and 2G.R. Anantha Raman

1P.G. Scholar, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India
2Assistant Professor, Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India

Received February 2016; Accepted 18 April 2016; Available 25 April 2016

ABSTRACT
Connectivity to the entire world has become extremely easy in the current situation. The only thing which makes connectivity look better is Big Data. Big Data is the one which enhances the quality of things in a smarter way through connectivity. It tries to manage and connect large clusters of data. Collecting and analyzing a review is a hectic task due to huge variety of data in different websites. This paper is an approach that classifies the review of hotels in Beijing. It easily communicates and tells the user of which is the best choice of hotel in Beijing through their selected query. It filters the unwanted data from the view of the user and provides the recommendation collaboratively to the user.

KEYWORDS: Big Data, Collaborative Filtering, Recommender Systems, POS, Manhattan Distance

INTRODUCTION
Today, people are so advanced that for everything they use Internet. Without Internet, nothing becomes possible with almost 85% of the people around the world. Starting from transactions till shoe purchase, everything has become online. For managing this vast amount of people online, we need some technology which will be good in coverage and also speed. With this idea, Big Data is used today in almost everything. Big Data finds to be good in coverage and also speed that will satisfy the needs of the user.

As said earlier, people are so advanced that even for purchasing, travel, bill payment or for cooking they use Internet. While speaking about purchase, a user always wishes to opt for the perfect product or a thing which is the best in its quality. And even to find a product and its quality, people use Internet. Only with this idea, this paper has been introduced where it provides the recommendation to the user of which hotel is the best one to choose in Beijing based on the query which they choose.

This is done through the review of different users in different websites. All the reviews are first collected from different websites. Then, through the process of POS Tagging, the sentences are split into words where it is pre-processed for finding the relevant data. Then the similarity between the terms is found through Manhattan distance. Based on the similarity between the terms, they are clustered together. From the cluster, the recommendation is given to the user based on the query.

The rest of the paper is organized as follows: In Section II, the methodology is detailed, in Section III, the results obtained through implementation are discussed, in Section IV, the performance of the system is evaluated and in Section V, few of the conclusions are drawn.

I. Related Work:
Jianying et al. [3] presented a neural network-based clustering and collaborative filtering algorithm where the cluster analysis gathers the users with similar characteristics according to the web visiting message data. Through examining the algorithm, it is inferred that it is hard to say that a user’s preference on web visiting is relevant to preference on purchasing. Mittal et al. [10] designed a Recommender System framework through K-means clustering algorithm which was applied to partition the movies based on the selection of the user. It is inferred that it requires users to provide some extra information in order to evaluate the choice of the user. Xiaohui et al. [16] came with an approach that used multi-dimensional clustering based collaborative filtering that provides a quality recommendation to the users. This approach was likely to trade-off on increasing the diversity of the recommendations. Zhangbing et al. [19] proposed a fuzzy C-means algorithm where the similar services are merged into same cluster by which the capacity of the service search engine was improved significantly. It is inferred that the domain ontology exists for facilitating semantic interoperability. Renaud et al. [20] combined both the collaborative filtering and clustering where it requires only implicit feedback on past user purchases to discover the relationship between the users. Based on the clustering results, the products of high interest were recommended to the users. The drawback here is that implicit feedback does not always provide sure information about the preference of the user.

**Architectural Design:**

![Fig. 1: Architectural Design.](image1)

The architectural design of the entire system is detailed in Fig. 1.

**II. State Of The Art:**

**A. Data Selection:**

Before starting any process, the first thing to do is the selection of data. Without Data selection, none of the process can be done. Here the review of hotels in Beijing is collected from different websites. The reviews selected here are done by the valuable customers who have already visited the hotel. The process is detailed in Fig. 2.

![Fig. 2: Review of different users.](image2)

The data is crawled first from different websites and then combined together to give the recommendation of the best hotels to the users.

**B. POS Tagging:**
Parts-of-speech Tagging is simply denoted as POS Tagging. POS is nothing but, it will assign to each word the part of speech, i.e. it will split the sentence into words by assigning POS to each sentence. The process of POS Tagging is detailed in Fig. 3.

Once POS Tagging is assigned to the sentence, it becomes easier to process even terabytes of data in few seconds. POS Tagging will split the sentence into words which does not create any meaning.

C. Pre-processing:

POS Tagging will never create meaning. To create meaning to the tagged words, the words need to be pre-processed. Pre-processing is discussed in Fig. 4

The pre-processed result will yield meaningful words which is the one helpful for calculating the frequency of the words and the documents.

D. Term Frequency:

The term frequencies need to be calculated for the pre-processed words. This process is carried out for finding the similarity between the terms occurring in a particular document. The process is shown in Fig. 5

The term frequency is calculated in such a way that it is a measure of number of the similar terms in the document to the total number of terms in the particular document. The dissimilar terms will have a measure of 0.0

E. Inverse Document Frequency:

Fig. 6 calculates the IDF value. IDF or Inverse Document frequency is helpful in determining the similarity between the documents. IDF is the measure of the number of similar documents to the total number of available documents.
F. Manhattan Distance:
To calculate the Manhattan Distance, the frequencies of the terms and the documents need to be computed together. Frequency sum is shown in Fig. 7.

Fig. 5: Calculating the Term Frequency.

Fig. 6: Calculating the Inverse Document Frequency.

Fig. 7: TF*IDF.

Once the value is computed, then the Manhattan Distance needs to be computed. Distance calculation is shown in Fig. 8.
The formula for computing the Manhattan distance is given in \( F.(1) \)

\[ X_i - Y_i \]

Where the terms \( x \) and \( y \) are the values of relevant documents.

**Fig. 8:** Manhattan Distance.

**G. Clustering:**

The similar values computed through the Manhattan Distance will be clustered together in this stage. The clustering process is done through the computation of the median value. Clustering process is shown in Fig. 9.

**Fig. 9:** Clustering.

Once the median value is calculated, the cluster is formed based on the relevant and similar values. Now from the selected query of the user, the best list of hotels in Beijing will be given to the user. This process is detailed in Fig. 10

**III. Performance Metrics:**

The success of any system which is proposed depends upon how fast the entire system works and how accurate it is. This means the computation time and accuracy of the system. How big is the system, the computation time of the processes being carried out should be minimum and the accuracy should be maximum.

To prove the minimum computation time and maximum accuracy of the system, the system is compared with another existing system and it is found that the computation time and accuracy of the proposed system is far better than that of the existing system.

As shown in Fig. 11, the Computation time of the proposed system is found better than the existing system. The accuracy of the proposed system is found better than the existing system as shown in Fig. 12.

**Conclusion:**
This paper effectively classifies the review of hotels in Beijing. More precisely, the recommendation is given to the user through the frequencies of the occurrence of terms and through the Manhattan Distance calculation by the way the computation time is minimized to the extreme and accuracy is increased highly. The collaborative recommendation is helpful to the user where it is more accurate as expected by him/her and also faster.

Fig. 10: Recommendation.

Fig. 11: Comparison of computation time.

Fig. 12: Comparison of accuracy.

REFERENCES