Open access Journal International Journal of Emerging Trends in Science and Technology

Collaborative Filtering Recommender System

Authors
Anvitha Hegde[1], Savitha K Shetty[2]
[1]M.Tech, Dept. of ISE, MSRIT, Bangalore
[2]Assistant Professor, Dept. of ISE, MSRIT, Bangalore

Abstract
There has been an increase in the number of services available in the internet .Datasets are growing at a fast pace as it is being gathered and generated by a number of devices like smartphones, tablets and various information sensing devices. Traditional data processing methods are ineffective in handling such a huge amount of data related to services within limited time constraints. Most of the present day recommendation systems use structured data. In order to handle the large amount of data relevant to the services and assist a user in selecting a service which is most relevant a collaborative filtering based recommender system is proposed which uses unstructured data. There are three stages in this method. In the first stage porters stemming algorithm is applied ,then clustering is applied on the data, in order to reduce the number of services, in the last stage a filtering approach is used in order to recommend relevant services to the user. As stemming and clustering are applied before filtering recommendations are done at a faster pace.
Keywords: *stemming, clustering, collaborative filtering, pearson coefficient, recommender system.*

INTRODUCTION

Big data refers to data which are complex and has more volume. Big data involves data from a wide variety of sources. The data can be structured, unstructured or a combination of both. Big data generally refers to petabytes and terabytes of data.Big data has managed to gather attention of various industries, academic institutions[1] and is also used in government funded projects. There is a need to deal with real time data as data is being generated constantly and changes very frequently. It is a necessary to apply a data processing and data mining techniques in order to extract useful information[2] from this huge amount of data. Traditional data mining tools fail to capture real time data within tolerable time limits[3]. A user searching for data has to deal with a potentially overwhelming set of data. In order to help a user in selecting relevant data recommender systems are used assist users in taking a decision and provide a set of alternatives which the user can choose from.

Recommender systems are systems which enables user to select the relevant items from a large number of items[4]. They help in extracting relevant data from an overwhelming amount of data. Usually filtering is done by extracting structured data (data stored in rows and columns).In the recommender system being discussed here unstructured data has been used.

DIFFERENT TYPES OF RECOMMENDER SYSTEMS

Hybrid Recommender System

These types of recommender systems[5,6] combine two or more form of recommender systems. Several hybrid methods have been developed which combines content-based filtering and collaborative filtering approach or demographic filtering and content based filtering approach.

These recommender systems can overcome the disadvantages of some of the most common recommender systems. The results of two types of recommender system can be combined to give a recommendation or the input of a recommender system can be given to another recommender system.

Advantages

- The overall accuracy of this combination of recommender systems is improved and content-based filtering could be more effective in some cases.
- Hybrid recommender systems overcome the problem of sparse matrices.

Disadvantages

- Complex to develop as it is a combination of two or more recommender systems.
- It is not very effective when the number of users is small
- Time consuming as two types of recommender systems has to be run parallel.
- Cannot be applied to unstructured data.

Demographic filtering recommender systems

In this type of recommender system[7] the locality information of users is used. Thepreferences of users who belong to the same locality are extracted and the most popular items among the extracted preferences are recommended to the user.

Disadvantages

- Intrusive in nature as locality information is gathered.
- Information of the location of the users is not always available.
- Items cannot be predicted if there are not many users who belong to the same location as that of the user.

Knowledge based recommender system

The information about the relationship[7] of the user with the item is gathered and the items are recommended based on the relationship.

Disadvantages

- Domain knowledge of the services or products is necessary.
- History of the browsing habits of the user must be available in order to make the appropriate recommendations to the user.

Collaborative recommendation system

These recommender systems[8,9] make recommendations to the user using the preferences of other users and by using the preferences of that particular user to the various items.

Advantages of collaborative filtering recommender system[10]

- The items are not represented in terms of features and attributes.
- The most popular items are recommended.
- Domain knowledge is not necessary.
- Recommendations are improved over time.

Disadvantages:

- If there are few users who have preferred an item then the item won't be recommended to others.

IMPLEMENTATION

Unstructured data[11] refers to the data which doesn't have a predefined structure. The unstructured data is converted into a json format which has the following components<content,id ,title, segment,boost,digest, stamp,description>.

The components of the data are as follows

- Content: Information about the contents and functionality provided by the service
- Id:url of the service
- Title: the title of the dataset.

- Segment: the segment number of the dataset.
- Boost: time required to fetch the dataset.
- Digest: A unique number generated for each dataset.
- Tstamp: time and date of the creation of the service.
- Description: A brief description of the service.

Specification of the collaborative filtering approach

Stage 1:

Step 1.1: convert the unstructured dataset into json format

Step 1.2: apply porters stemmer [12]

Step 1.3: compute description and content similarity using jaccard similarity coefficient.

Step 1.4: compute characteristic similarity by taking the sum of the description and functionality similarity. Create a matrix M each element of which is characteristic similarity.

Step1. 5: Cluster services according to the characteristic similarity in M using agglomerative hierarchical clustering[13].

Stage 2:

Step 2.1: compute rating similarity using pearson coefficient[14]

Step 2.2: select the services whose rating similarity exceeds a given threshold. And put them into the neighborhood set.
Step2. 3: recommend the service to the user if the rating exceeds a given threshold.

Stage 1

Stemming

Stemming[15] is a process in which different grammatical variations or deviations of a word are mapped to the root word. Stemming is used in query processing applications .while processing a query the stemming algorithm usually reduces the number of documents retrieved as it maps several terms to a single word. Thus accurate documents are displayed to the user. In the simplest form of stemming, all the root words are stored in a table and the table is queried to find a word which matched with the given word. Some algorithm removes the suffixes or affixes (prefix and suffix) before searching for the matching words. In certain stemming algorithms the suffixes are substituted then the matching words are searched. A stem dictionary is maintained and the root word id searched in the dictionary in some algorithms.

Step 1: measure the number of consonants and the number of vowels in a word

Step 2:Get rid of the plurals present in the word. Example:ed,ing etc.

Step 3: Convert the letter y to i if there is another vowel present in the word

Step 4: Convert the double suffixes present in the word to a singular form. (ization is transformed to ize).

Step 5: Remove the suffixes less and full from the word.

Step 6: Strip ant and ence suffixes present in a word.

Step 7: remove e from the word.

Step 8: extract the root form of the word.

The above stemming algorithm is applied to the content and description field of the dataset and the description and content similarity is obtained using jaccard similarity and then the characteristic similarity is obtained by using the description and content similarity. The characteristic similarity matrix is used as the input in the clustering stage.

Clustering

Input: the set of services, characteristic similarity matrix (M).
Output:dendrogram
Step 1:let the initial level be L(0) = 0 .All the data points are in separate clusters.

Step 2: Find the distance between neighbouring clusters in the current clustering, say pair (a), (b), according to d[(a),(b)] = min d[(i),(j)] where the minimum corresponds to the least distance over all the clusters.

Step 3:Merge clusters (a) and (b) into a single cluster to form the next clustering m. Set the level of this clustering to d[(a),(b)].

Step 4: Update the distance of the clusters as and when they are merged.

Step 5: Stop when all the data points are in a single cluster**.**
The resulting set of services are used as the input to the second stage.

Stage 2

Step 1: Create a model of the input datasets (which contains the ratings).

Step 2:compute thepearson correlation similarity to obtain a predicted rating.

Step 3:Set a threshold for the neighbourhood.

Step 4:Extract the services which is above the threshold.

Step 5:recommend the services.

if the predicted rating of a service exceeds a recommending threshold, it will be a recommendable service for the active user. A service is generally rated on a positive point scale from 1 (very dissatisfied) to 5 (very satisfied). Therefore, the recommending threshold is set to 2.5 which is the median value of the max rating. All recommendable services are ranked in non-ascending order according to their predicted ratings so that users may discover valuable services quickly.

TIME COMPLEXITY ANALYSIS

The time complexity of the recommender systems is divided into two parts. The time required for stage 1 (stemming and clustering) and time required for stage 2(collaborative filtering).
In the agglomerative hierarchical clustering the first step is to compute the similarity between every pair of services. If there are n number of services the time complexity[16] for this step is $O(n^2)$.In the second step the pairs of most similar clusters are selected. In the initial step all data points are in a separate cluster later they are merged based on the similarity, this step has the complexity $O(n-l)^2$ .The insertion and deletion operations take O(n-l)log(n-l) time. The overall complexity is approximately $O(n^2 \log n)$.
If there are i number of users and j number of services and the relationship between then is denoted by a i*j matrix, the time complexity to determine similarity using Pearson similarity measure is $O(ij^2)$ As clustering is applied before the filtering stage, the number of services are reduced. If i_k is the reduced set of services and j_k is the number of users who have rated that service, the time complexity becomes $O(i_k j_k^2)$. As $i_k < i$ and j_k is less than j the time complexity of this approach is reduced.

EXPERIMENTAL SETUP

Data from programmableweb website has been used. The datasets is composed of APIs. (application programmable interfaces).The recommender system is used to predict the APIs.Apache nutch was used to craw the website and gather the dataset. The dataset was stored in apache solr.The collaborative filtering layer was deployed on apache mahut.

CONCLUSION AND FUTURE WORK

The recommender system is composed of three stages, stemming, clustering and recommendation. Recommendations can be done at a faster pace as the data is first clustered then the filtering

technique is applied. Most of the recommender systems are applied on structured data where data is stored in rows and columns however this system is applied on unstructured data. The time consumed is very less in this system as filtering is done on the clustered set of services which has reduced number of services and ratings of the clustered services are more relevant to each other compared to that of dissimilar services.

In the future the recommender system can be enhanced by gathering information from the bookmarks and by using the history of the users browsing habits. The demographic information can be extracted and used in the recommender system, as the information about a user's location can be very helpful in making appropriate recommendations, for example a user searching for a hotel can be presented with a list of hotels which are in the same locality as that of the user.

REFERENCES

1. M. A. Beyer and D. Laney, ``The importance of `big data', 'Gartner Inc., Stamford, CT, USA, Tech. Rep, 2012
2. X. Wu, X. Zhu, G. Q. Wu, and W. Ding, ``Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97107, Jan. 2014.
3. A. Rajaraman and J. D. Ullman, Mining of Massive Datasets. Cambridge, U.K.: Cambridge Univ. Press, 2012
4. A. Bellogín, I. Cantador, F. Díez, P. Castells, and E. Chavarriaga, `an empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. Intel. Syst. Technol., vol. 4, no. 1,Jan. 2013.
5. Hybrid Recommender Systems: Survey and Experiments† Robin Burke California State University, Fullerton,2013.
6. A. Bellogín, I. Cantador, F. Díez, P. Castells, and E. Chavarriaga, 'an empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. Intel. Syst. Technol., vol. 4, no. 1, Jan. 2013.
7. Applicability of Demographic Recommender System to Tourist Attractions: A Case Study on Trip Advisor**,** Yuan yuan Wang ; Dept. of Compute., Hong Kong Polytechnic. Univ., Hong Kong, China ,2014.
8. Using Content-Based Filtering for Recommendation, Robin van Meteren1 and Maarten van Someren2,Jan 2013.
9. Collaborative Filtering Recommender Systems by Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, June 2014.
10. Evaluating collaborative filtering recommender systems, Jonathan L Herlocker, John T.Riedl, and June 2012.
11. Z. Zheng, J. Zhu, and M. R. Lye, ``Service-generated big data and big data as-a-service: An overview," in Proc. IEEE Int. Conger. Big Data, Oct. 2013
12. Porter M.F. "An algorithm for suffix stripping". Program. 1980; 14, 130-137.
13. Porter M.F. "Snowball: A language for stemming algorithms". 2001.
14. Y. Zhao, G. Karis, and U. Fayyad, ``Hierarchical clustering algorithms for document datasets," Data Mining Knowl. Discovery, vol. 10, no. 2,Nov. 2005.
15. Mai, Y. Fan, and Y. Shen, ``A neural networks-based clustering collaborative filtering algorithm in e-commerce recommendation system," in Proc. Int. Conf. Web Inf. Syst. Mining, pp. 616619, Jun. 2013.
16. J. Wu, L. Chen, Y. Feng, Z. Zheng, M. C. Zhou, and Z. Wu,``Predicting quality of service for selection by neighbourhood-based collaborative filtering," IEEE Trans. Syst., Man, Cyber., Syst., vol. 43, no. 2,Mar. 2013