Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

Keyword based Movie Recommendation Service using MapReduce

Asiya Banu B.
Senior Grade Lecturer,
Department of Computer Science and Engineering
Government Polytechnic,
Harihara, Karnataka, India.

Shaheen Banu.
Senior Grade Lecturer,
Department of Computer Science and Engineering
D.R.R. Government Polytechnic,
Davanagere, Karnataka, India.

***Abstract*— Recommender Systems provide customers with information in the form of recommendations that helps to decide which products to purchase. In recent days these traditional Recommender Systems are facing a service overload problems as the data is increasing tremendously and analyzing such large dataset is a challenging task. Such large volume of data is termed as Big Data. Most of the existing service Recommender Systems provides same rating to users without considering diverse user's preferences. User based Collaborative Filtering approach for recommending appropriate service to users based on their preferences has been proposed. The model uses weighted average approach to calculate the personalized rating of a service for the active user. The proposed model will mainly focus on custom-made service recommendations which use keywords to indicate current user-preferences and the model is implemented in HADOOP to meet its scalability and efficiency.**

Keywords—Recommender Systems, Big Data, Collaborative Filtering, HADOOP.

I. INTRODUCTION

"The amount of information in the world is increasing exponentially than our ability to process it [1]. The challenges for data include capture, analysis, curation, search, storage, sharing, visualization, transfer, and privacy violations.

Recommender Systems are adaptive systems that deliver suggestions to their users about the content that matches their estimated interests. Two major strategies are content-based and collaborative recommendations. Content based systems recommend those items that resemble the ones the user liked in the past, while collaborative systems recommend the items that the other users with similar preferences liked in the past [4].

II. RELATED WORK

There have been number of Recommender Systems proposed both in academia and industry. In [1], the authors introduced Recommender systems which applied statistical and Knowledge Discovery techniques to the product recommendation to live customers. In particular traditional data mining algorithms, nearest neighbor collaborative filtering and dimensionality reduction algorithms were applied on two different varieties of data sets. In [3], the authors proposed Item-to- Item Collaborative Recommendation algorithm which can scale to massively huge data sets and generate superior quality recommendations in real time environment. The algorithm matches similar items to rated and purchased items then a recommendation list is generated by combining those similar items.

III. MOTIVATION

Existing Service Recommendation Systems suffer from scalability and are inefficient in analyzing and processing Big Data and same set of user ratings are displayed to different users without considering the variety of preferences of users, hence fails to achieve/meet user's personalized requirements. The proposed system aims at presenting a customized service recommendation list by recommending the most appropriate services to the users. A User based Collaborative Filtering technique is adopted to create a list of appropriate recommendations. The similarity in preference between current user and past users is calculated based on the similarity in the rating history of the past users by approximate similarity computation. Weighted average approach is used to calculate customized rating for the active user.

IV. PRILIMINARY KNOWLEDGE

A. Big Data

In a Big Data environment the size of the datasets is so large that it cannot be captured, stored and managed and analyzed by using any traditional software tools [5]. Big Data management emerged out as a challenge for IT companies. Service recommender systems are one of the valuable tools that help users to deal with service overload and provide appropriate recommendations to users and analyzing such large volume of data sets is called "Big Data" [7].

B. MapReduce

MapReduce is a software framework and programming model for large-scale distributed computing on massively huge amount of data. The MapReduce framework implementation was adopted by an Apache Software Foundation and named it as Hadoop. The two phases MapReduce framework are the map phase and the reduce phase.

The processes can be specified by the below two functions:

1. In the Map phase, the map functions are executed in parallel with various input splits which is stored in a local distributed file system named Hadoop Distributed File System (HDFS). All intermediate values related with the same intermediate key are grouped together by the MapReduce library and passes them to the reduce function.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

2. In the reduce phase, the intermediate key are accepted along with the set of values and then merges these values together to generate probably smaller set of values.

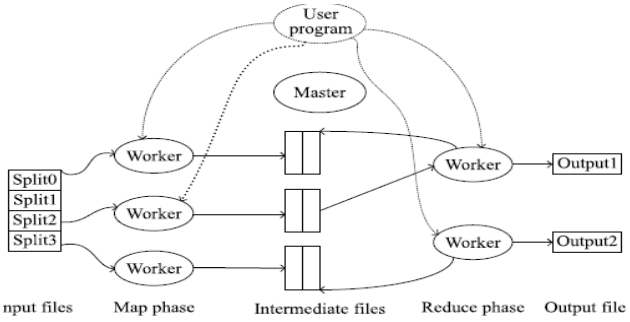

Fig 1:The Implementation of MapReduce Model

Figure 1 illustrates the map phase, in which the input data is first split called as input_splits and those input_splits are then feed to workers. Each Individual data item is called a record which is further read by RecordReader. The input_splits from each worker are parsed by the MapReduce system to produce the records. Intermediate results generated in the map phase are then shuffled and sorted by the MapReduce system in the reduce phase. Multiple reducers compute final results and write it to the disk [8].

C. Domain Thesaurus

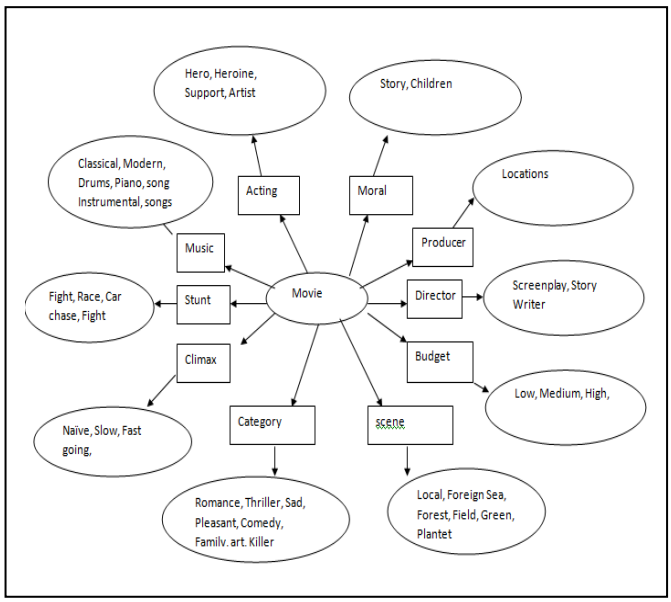

Figure 2: Simple Domain thesaurus of a Movie Recommendation Service.

A Simple example of Domain thesaurus for a Movie Recommendation service is shown in figure 2. The keywords are formed into corresponding Keyword candidate list as shown in the below table 1. Words in the reviews are written by the previous users may not exactly match the corresponding keywords in the keyword candidate list but may reflect the same aspect; such words should be extracted as well. The words in the rectangular shaped boxes refer to the keyword candidate list. The words in the oval shaped boxes refer to the domain thesaurus.

Table 1: Keyword candidate list

No.	Keyword	No.	Keyword
1	Acting	6	Scene
2	Moral	7	Category
3	Producer	8	Climax
4	Director	9	Stunt
5	Budget	10	Music

III. SYSTEM ARCHITECTURE

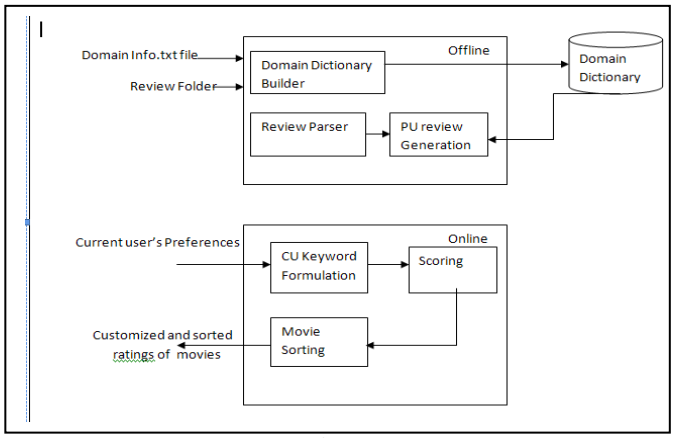

Figure 2: System architecture of Movie Recommendation Service

Figure 2 depicts the different modules of the system. The system consists of Domain Dictionary Builder, Review Parser, PU review Generation which will be performed during offline. The online parts are CU keyword formulation, Scoring and finally Movie Sorting.

The Module Description proposed System Architecture is

A. Domain Dictionary Builder

Keywords are used to define users' preferences of candidate services. The set of keywords define users' preferences of the candidate services, which can be denoted as, $\{k_1,k_2,k_3\ldots\ldots k_l\}$ *l* is the number of the key-words in the keyword-candidate list. Domain Info text file is given as input to Domain Dictionary Builder from which keywords are extracted to store it in Domain Dictionary.

B. Review Parser

Preferences of past users preferences will be extracted from the Reviews Folder and formalized into a keyword set. Usually, since the words written by the past users were in natural language and those reviews need not exactly match the corresponding keywords in the keyword-candidate list but may refer to the same aspects as the keywords. The corresponding keywords should be extracted as well. Specialized domain thesauruses are built to support the keyword extraction and PU is generated.

C. CU Generation

A current user can give his/her preferences about candidate services by selecting keywords from a list of keywords. The preference keyword set of the active user can be de noted as CU = $\{ck_1,ck_2,\ldots\ldots ck_l\}$ where ck_i $(1\leq i\leq l)$ is the i^{th} keyword selected from the keyword-candidate list by the active user, *l* is the number of selected keywords.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

a) Scoring

User based collaborative filtering is used to identify users who have similar tastes with the active user.

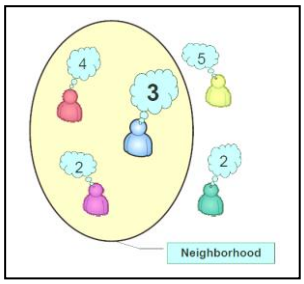

Figure 3: User Based Collaborative Filtering

Approximate similarity computation is used to group past users who have similar preferences to current users by finding neighborhoods of the current user. Figure 4.3 explains User based Collaborative Filtering approach, a frequently used method for comparing the similarity and diversity of sample sets, Jaccard coefficient, is applied in the approximate similarity computation.

$$Sim(CU,PU)= Jaccard\ (CU,PU)= \frac{|CU \cap PU|}{|CU\ U\ PU|}$$

Where, CU refers to the preference keyword-set of the active user, PU is the preference keyword-set of a previous user.

D. Movie Sorting and generating recommendations.

1Once the set of most similar users are found, the personalized ratings of each candidate service for the active user can be calculated. Finally, a personalized service recommendation list will be presented to the user and the service(s) with the highest rating(s) will be recommended to him/her.

Basic Algorithm for Recommendation Service

Input: The preference keyword-set of the active user APK.
The number K,the threshold in the filtering phase.
Output: The services with the Top-K highest ratings.

1: for each service ws_i belongs WS
2: Ř=ø, sum=0, r=0
3: for each review R_j of service ws_i
4: process the review into a preference keyword set PU_j
5: if $PU_j \cap CU \neq ø$ then
6: insert CU_j into Ř
7: end if
8: end for
9: for each keyword set PPK_j belongs to Ř
10: $sim(CU,PU_j)$=IDENTICAL(CU, PU_j)
11: if $sim(CU, PU_j) < \partial$ then
12: remove PU_j from Ř
13: else sum=sum+1,r=r+r_j
14: end if
15: end for
16: $\bar{r}$=r/sum
17: get pr_i by weighted average approach.
18: end for
19: sort the services according to the personalized ratings pr_i
20: return the services with the Top-K highest ratings.

D. Tables

TABLE I. SYMBOLS AND ITS DEFINITIONS

Symbols	Definition
{ k_1,k_2,k_3…… k_n }	No. of key words in the keyword candidate list
CU	Current user's keyword set
PU	Past user's keyword set
$Sim(CU,PU_j)$	Similarity between current and Previous users.
$\bar{r}$	the average ratings of the candidate service.
Pr	Personalized rating of all the candidate services for the Active user.

V. CONCLUSION

The proposed system uses keywords to specify current users' preferences, and a User Based Collaborative Filtering algorithm is adopted to produce appropriate recommendations. Domain thesaurus is used to identify and extract users' preferences. The current user gives his/her preferences in the form of a text file, which consists of keywords from keyword candidate list. The preferences of the past users are extracted from reviews for services according to the keyword-candidate list and domain thesaurus. Neighbourhoods of the current user are found and the reviews unrelated to the current user's preferences will be filtered out by the intersection concept in set theory. Weighted average approach is used produce personalized recommendations recommending the most suitable service(s) to the users. To improve the scalability and accuracy in "Big Data" environment, it is implemented on a MapReduce framework in Hadoop platform. The proposed system significantly improves the scalability and accuracy of service recommender systems over existing approaches.

REFERENCES

[1] B.M. Sarwar et al., "Item-Based Collaborative Filtering Recommendation Algorithms," 10th Int'l World Wide Web Conference, ACM Press, 2001, pp. 285-295.

[2] J. BEN SCHAFER et al., "E-Commerce Recommendation Applications", Data Mining and Knowledge Discovery, 5, 115–153, 2001

[3] G. Linden, B. Smith, and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, Vol. 7, No.1, pp. 76-80, 2003.

[4] M. Bjelica, "Towards TV Recommender System Experiments with User Modeling," IEEE Transactions on Consumer Electronics, Vol. 56, No.3, pp. 1763-1769, 2010.

[5] J. Manyika, M. Chui, B. Brown, et al, "Big Data: The next frontier for innovation, competition, and productivity," 2011

[6] Mukta kohar,Chhavi Rana: Survey Paper on Recommendation System.

[7] Shunmei Meng,Wanchun Dou, Xuyun Zhang, Jinjun Chen.KASR: A Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications.

[8] Yaxiong Zhao,Jie Wu, and Cong Liu, A Data Aware Caching for Big-Data Applications Using the MapReduce Framework.

[9] http://en.wikipedia.org/wiki/Big_data.

[10] Ian Sommer Ville. Software Engineering. 3rd edition, Pearson Education, 2011.