International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Recommendation of Services By Mapping Categorical Preferences With Past User Comments Using Hadoop

Ms.Ruchita V. Tatiya[1], Prof. Archana S. Vaidya[2]

[1]*Savitribai Phule Pune University, Computer Engg.Dept.,Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies & Research, Nashik, India, ruchitatatiya@gmail.com*
[2]*Savitribai Phule Pune University, Computer Engg.Dept.,Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies & Research, Nashik, India,archana.s.vaidya@gmail.com*

Abstract—Recommendation of services attempts to give the guidance to the users regarding the useful services based on their personalized preferences, past behavior or their similar likings with the other users. The various approaches of recommendation systems, like content-based, collaborative filtering, hybrid, etc, can further be classified according to their algorithmic technique as memory-based (heuristic) or model-based recommendation algorithms. Ratings and rankings are only being considered by many existing recommendation systems, ignoring the categorical preferences and likings of the users. Big Data, a major issue, is contributing to a large amount of data which is not only difficult to capture, store, manage and analyze but is also affecting many recommender systems with inefficiency and scalability problems. Proposed system enhances existing recommendation systems to generate recommendations based on the categorical preferences of the present user by matching them with the reviews/comments of the past users'. Keywords are used to indicate the preferences of present and past users and similarity between them is computed using methods like approximate and exact similarity computation. Semantic analysis is performed on the reviews to eliminate the unnecessary reviews of the users. To improve the performance of recommendation system in big data environment it can be ported on distributed computing platform, Hadoop which uses a Map-reduce computing framework.

Keywords-Service recommender systems; Big Data; Semantic analysis; Similarity computation algorithms; Hadoop; Map-Reduce.

I. INTRODUCTION

Information filtering systems removes the unnecessary information before presenting it to the human user. The subclass of these systems, called as recommendation systems helps the users to predict the services or items that the user would like. Service recommender systems [1] provide apt recommendations of the services to the users and have become popular in variety of practical applications like recommending the users about hotels, news, books, movies, music, travel, and products in general. The increase in the number of Internet users is contributing to immense amount of data every day. Such immense data, known as Big-data is not only difficult to capture and store, but also managing, processing and analyzing such data with the available current technology within the tolerable speed and time is a difficult task. This Big-data management also poses a heavy impact on service recommender systems.

II. MOTIVATION

The service recommender systems present the same ratings and rankings of the services to the different users and also provide the same recommendations to them without considering the user's personal likings and taste. Therefore these recommending systems do not fulfill the need of personalized requirements for the users [1]. Also many recommendation systems provides single-criteria ratings i.e. just the overall rating of any service is being considered which makes them less accurate [2]. To avoid this, multiple-criteria rating i.e. ratings given to the individual parameters of any service must be incorporated while generating the recommendation [3]. Due to the ever increasing amount of data, the Big-data management poses a heavy impact on service recommender systems with issues like scalability and inefficiency. The proposed system considers the issues and drawbacks of the existing system and contributes to generate recommendations more accurate according to the user likings and the categorical preferences and also tries to improve the efficiency and performance of the system in the big data environment.

III. LITERATURE SURVEY

There are various recommendation methods based on the information or knowledge source they use for making the apt recommendations. Reference [4] describes various methods to generate recommendations and also focuses on the algorithmic methods like memory-based and model-based algorithms. Author Manisha Hiralall [5] has compared various methodologies which can be used to generate recommendations for the varied web applications. Different pros and cons are also stated which helps the user to select the appropriate approach according to his/her application. Adomavicius and Tuzhilin [2], has described the current generation of the recommendation methods and have stated that they are based on rating and rankings only, without considering the taste and choices on an individual and just provide the recommendation based on the single criteria ratings. To overcome the drawback of single-criteria ratings, authors G. Adomavicius and Y. Kwon [3] incorporated and leveraged multi-criteria rating which improved the accuracy of the system as compared with single-rating recommendation approaches. A problem faced by many recommendation algorithms is its scalability, i.e. when the volume of the dataset is very large, the computation cost would be very high. The development of cloud computing software tools such as Apache Hadoop, Map-Reduce, and Mahout, made possible to design and implement scalable recommender systems in Big-data environment. Authors Z. D. Zhao and M. Shang [6] implemented the collaborative filtering algorithm on the cloud-computing platform, Hadoop which solves the scalability problem for large scale data by dividing the dataset. Shunmei Meng, Wanchun Dou, Xuyun Zhang and Jinjun Chen [1] proposed a keyword aware service recommendation method, which utilizes the reviews of previous users to get both, user preferences and the quality of multiple criteria of candidate services, and computes similarity with the preferences of active user which in turn makes the recommendations more accurate. Moreover, they implemented their approach on Map-Reduce which showed favorable scalability and efficiency. Peter D. Turney[7] presented an algorithm for classifying the reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. The algorithm presented has three steps: extract phrases containing adjectives or adverbs, estimate the semantic orientation of each phrase, and classify the review based on the average semantic orientation of the phrases.

IV. PROPOSED SYSTEM

4.1Problem Definition

"To develop a system that can semantically analyze the past users' reviews and performs similarity computations between the categorical preferences of the present and past users to provide

personalized recommendation list of the associated services. To deal with and manage the large amount of dataset, the system can be implemented on distributed computing platform called Hadoop, which uses Map-Reduce computing platform".

4.2 System Architecture

The proposed recommender system is specially designed for the large scale data. While recommending particular service, the system mainly considers the user preferences and uses the previous users' comments/reviews which accounts to the immense data on the web. Below Fig. 1 shows the architecture of the proposed system, which is specifically the information filtering architecture which uses the distributed computing platform to reduce the processing time. Here we need to filter the previous users comments according to the active user preferences and also need to semantically analyze them for removing the negative reviews, to present a personalized service recommendation list. System manages to deal with large scale data with the help of Hadoop (a distributed computing platform using the Map Reduce parallel processing paradigm for big data). The processing of data can be distributed across various nodes by splitting the input into multiple Map() and Reduce() phases and the response time of the system can be decreased. To test the working of the system, test dataset regarding hotels is used that helps us to analyze the throughput of the system. Later a more generalized form of this system can be developed using precision of experiments.

4.3 System Flow

Before mentioning the system flow, following is the description of terminologies used into it.

- Aspect keyword list (AKL): It is a keywords set related to the users' preferences searching for a particular aspect and also multiple criteria regarding that service are mentioned into it. For e.g. if the service is recommending the hotels then the aspect keyword list will contain all the related keywords regarding the hotels like service, food, location, comfort, etc. [1].
- Thesaurus : A thesaurus is the group of words collected according to their similarity of the meaning. Basically a domain thesaurus is associated with the aspect keyword list and it consists of the words grouped together according to the similarity with the AKL, including related and contrasting words [1]. Also positive and negative words thesaurus consists of all the positive and negative words, phrases which are used in common natural language.

Following is the system flow which is divided into two parallel executable processes. Fig. 2 depicts the flow of the system diagrammatically and is explained below:

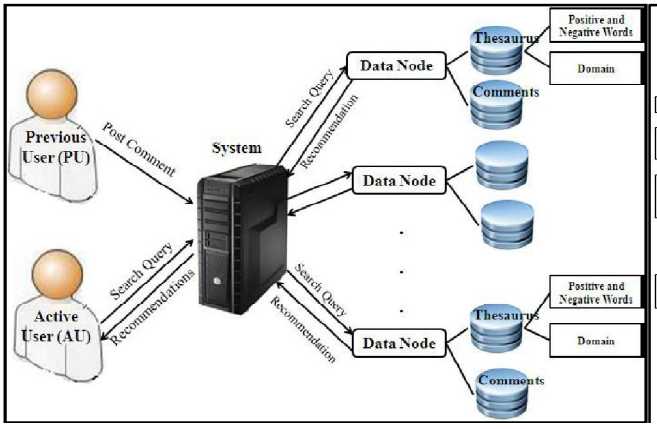

Figure 1.System architecture

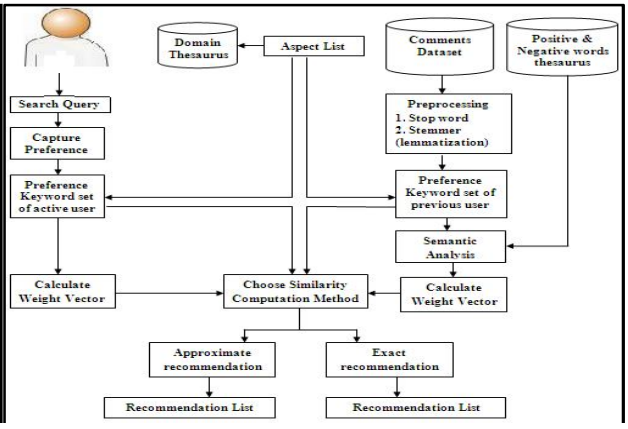

Figure 2. System flow

- Process 1 (For active user)

1) To get proper recommendation using the system, active user gives his/her preferences about the aspect of the services by selecting the keywords from an aspect keyword list, which reflect the quality criteria of the services he/she is concerned about. Besides, the active user should also select the importance degree of the keywords.
2) Finally, preference keyword set of active user and its weight vector is calculated using the AKL.

- Process 2 (For similarity computation)

1) Access the dataset having previous comments given by the past user.
2) Apply pre-processing like stop word removal, stemming and lemmatization
3) Using the keyword extraction technique calculate preference keyword set of previous user using domain thesaurus and the aspect keyword list.
4) Semantic analysis is performed on the preference keywords of previous user and the negative reviews are removed.
5) Calculate the weight vector for previous user preference keywords set.
6) Similarity computation – It identifies the comments of previous users whose taste matches to an active user by identifying the neighbors of the active user based on the similarity of their likings.
7) Calculate Approximate Similarity or Exact Similarity according to the user choice.
8) Generate Recommendation - Based on the similarity of the active user and previous users further filtering will be conducted and apt recommendations will be presented to the user if the similarity value is greater than the user defined threshold value.

V. IMPLEMENTATION

5.1 Environment

The proposed system is designed for open source operating system Linux - Ubuntu 14.04. The implementation of this system is based on Java jdk-7 and Hadoop 2.3 platform using the MapReduce framework. MySQL 5.5.41database is used for storing the datasets by configuring the LAMP server in Ubuntu. Also the configuration of php MyAdminin Ubuntu helps to perform various tasks such as creating, modifying or deleting databases with the use of a web browser. Eclipse (Luna) environment is being used for the system development. For recommendation generation latest version of Apache Mahout0.9 is used. To configure Apache Mahout 0.9 with Eclipse environment the integration of Maven 3.0.5 in Eclipse is done. To make the Mahout environment scalable it is being integrated with Hadoop. Initially for the testing purpose a single node Hadoop framework is being established.

5.2 Dataset

For the previous users comments or reviews regarding hotels, entity-ranking-dataset [8][9] is being used which is in the text format and contains: Full reviews of hotels in 10 different cities and there are about 80-700 hotels in each city which accounts to ~259,000 total number of reviews. The review format is : Date1<tab>Review title1<tab>Full review1. For creating Domain Thesaurus related to aspect keyword list, the use of Feature Words is done, downloaded from the Trip advisor (www.tripadvisor.com) site and was in the text format having the following form of : #cat=<category or aspect>. For semantic analysis of comments there is a need of positive and negative words. It has been downloaded from[10][11]. These lists of words were downloaded in the .xls format.

VI. RESULTS

6.1 Preprocessing Dataset

The datasets used for the system are in raw format and immense in nature which therefore requires huge processing to convert it in the usable format. The text and .xls files were converted into .csv format for further processing which were then imported into the MySQL database. Following fig. 3 shows the database created after this processing. The "catword" table stores the categorical preferences related to the hotel dataset and its related domain thesaurus words. The "positivekwd" and "negativekwd" table holds the positive and negative words which are used for the semantic analysis. The "review" table is created to store the pre-processed comments after applying the stemming and stop word removal algorithms on it, consisting of the following fields: comment id, file name, date, original review, stemmed review, stop-word removed review. Fig. 4 shows how the pre-processed comments are stored in the review table.

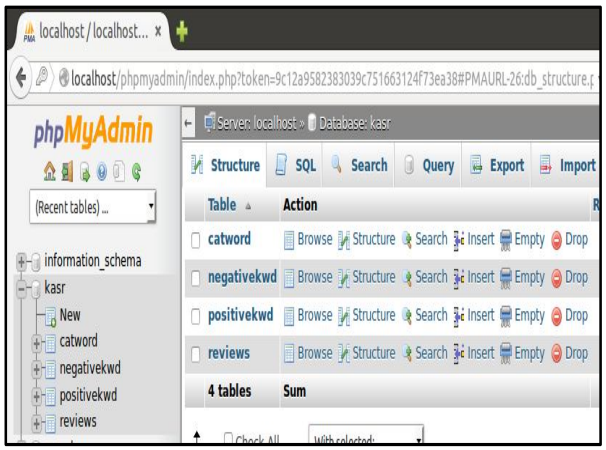

Figure 3.Database

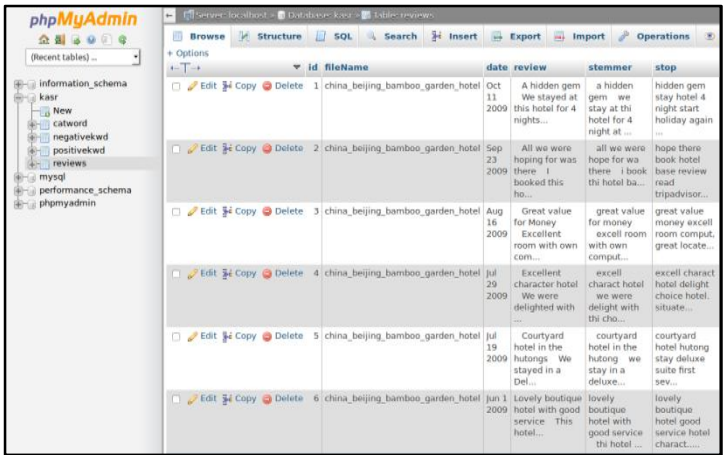

Figure 4. "Review" table

6.2 Semantic Analysis of Previous User Comments

The semantic analysis of the pre-processed comments was done to find out the negative words related to a particular categorical preferences so that the comments with the negative semantics should not be considered while generating the recommendations. Figure 5 shows the report generated after applying the semantic analysis on the preprocessed comment. The report consist of comment id, negative words, domain related to negative words and domains related to positive words.

CommentId	N Word	N Domain	P Domain
1	cheap problem	service	front_desk check_in front_desk chec...
2	cheap pleas problem	front_desk check_in location service	service business_service rooms
3			value rooms location location servic...
4			rooms
5	sever muddy	location	location location
6	problem complaint prohibit	service service	service rooms business_service
7			business_service
8			
9	forbidden		service
10			
11	dark		rooms service
12	trouble		
13	hard	location location	rooms front_desk check_in
14	pleas	location location rooms rooms	location service location
15	problem sever		location
16			rooms location rooms
17	hang	location	
18	pricey	service	rooms location
19			

Figure 5. Report generated after semantic analysis of previous users comments

6.3 Expected Results

The proposed system is expected to have a comparative analysis of the recommendation system with and without semantic analysis. Also the analytical reports will be generated consisting of accuracy and performance evaluation. Accuracy evaluation will be done for the approximate and exact recommendation system by using the F-measure. The response time and the scalability of the system can be analyzed by judging the performance of the system.

CONCLUSION

The keywords from the comments or reviews of previous users are used to indicate the preferences of the active users. An Aspect keyword list and Domain thesaurus are provided to help obtain users' preferences. The focus is on presenting a personalized service recommendation list to the users. The negative reviews of the users, can be avoided as the comments are semantically analyzed to increase the accuracy of the recommendations. As this accounts a large dataset, it is affected by the factors like scalability and inefficiency which can be improved by implementing the system in distributed cloud computing platform known as Hadoop which uses Map-Reduce framework and can manage large amount of data in these service recommendation systems.

REFERENCES

[1] ShunmeiMeng, Wanchun Dou, Xuyun Zhang and Jinjun Chen, "KASR:A Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications", IEEE Transactions On Parallel And Distributed Systems, vol. 25, no. 12, december 2014.
[2] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems : A Survey of the State of the Art & Possible Extensions", IEEE Transactions on Knowledge and Data Engineering,Vol.17,No.6 pp. 734-749, 2005.
[3] G. Adomavicius and Y. Kwon, "New Recommendation Techniques for Multicriteria Rating Systems", IEEE Intelligent Systems, vol. 22, no. 3, pp. 48-55, May/June 2007.
[4] Ruchita V. Tatiya and Prof. Archana S. Vaidya, "A Survey of Recommendation Algorithms", International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 6, Ver. V, PP 16-19, Nov – Dec. 2014.
[5] ManishaHiralall, "Recommender systems for e-shops", VrijeUniversiteit, Amsterdam, 2011.
[6] Z. D. Zhao, and M. S. Shang, "User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop", In the third International Workshop on Knowledge Discovery and Data Mining, pp. 478-481, 2010.
[7] Peter Turney,"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews".
[8] KavitaGanesan and ChengXiangZhai, "Opinion-Based Entity Ranking", Information Retrieval, 2011.
[9] For comments dataset : http://www.kavita-ganesan.com/entity-ranking-data
[10] For semantic analysis : http://mpqa.cs.pitt.edu/lexicons/effect lexicon/
[11] For semantic analysis, positive and negative words thesaurus : http://sentiwordnet.isti.cnr.it/download.php

