IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol.6, No.2, Mar-April 2016

Hotel Recommendation System using Hadoop and MapReduce for Big Data

Khushboo Ramesh Shrote
Computer Science and Engineering Department
Government College of Engineering Amravati, India
khushbooshrote166@gmail.com

Prof. Anil V. Deorankar
Computer Science and Engineering Department
Government College of Engineering Amravati, India
avdeorankar@gmail.com

***Abstract*—Success of web 2.0 brings online information overload. Google search engine is the basic motivation behind personalized recommendation. Google recommends services according to popularity of web page, number of hits or cache maintained by the end user. Most famous recommendation site tripadvisor.com also works on reviews of passive users. But the drawback is it only matches the number of reviews in which the keywords typed by an end user matches. No sentiment analysis is applied for ranking purpose. In this paper Hadoop framework based hotel recommendation system is proposed. Sentiment analysis is applied for score calculation. This total score will then used for recommendation purpose. Efficiency and Scalability is increased using Hadoop framework.**

Keywords- Keywords, Big Data, Hadoop, MapReduce, Sentiment Analysis.

I. INTRODUCTION

A personalized product/service recommendation isn't based on an assumption or guess. Personalized recommendations are based on user behavior. These are items that have been frequently viewed, considered, or purchased with the one the customer is currently considering. The personalized recommendation is based on large amounts of historical user data. This data must be in the form of Big Data. Because recommendations are more accurate in large amount of data. This data is often unstructured or semi structured in the form. Traditional recommender systems cannot fetch and examine such huge data. So, we use hadoop framework to deal with big data analysis problem.

Reviews are the source for personalized recommendation. In this paper, sentiment analysis method is applied to calculate total score corresponding to the reviews.

II. MOTIVATION

In almost traditional recommendation systems, such as tripadvisor.com matching keywords are found and the number of reviews matching with the users entered query is considered. No further calculation for reviews is done. So accuracy of result is low. Likewise in justdial.com local search engine recommendation is done by popularity of service and rating value. But the rating value not always be true. Rating can be increased by paying more money for the website.

Only passive user's reviews are most reliable. Because, users who visit the hotel or use any service can give his real opinion about that service. Sentiment analysis can give accurate results by using these reviews.

III. LITERATURE REVIEW

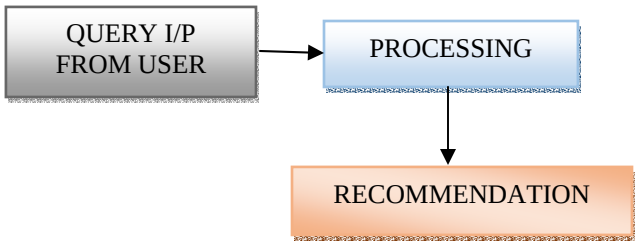
Following papers are reffered to study the progress in this area. Merits and Demerits are studied and a new concept is established.

Technology	Context	Merit	Demerit	Application
KASR: A Keyword-Aware Service Recommendation Method on Map Reduce for Big Data Applications	user based collaborative filtering algorithm is used. To make the method more efficient and scalable it is implemented on Hadoop. Jaccard coefficient and Cosine similarity measure is used for evaluation.	1. Scalable 2. More efficient than traditional methods	Jaccard Coefficient method is not so accurate. Users positive and negative reviews are not differentiated. Sentiments in the text is not considered for calculation.	Service Recommendation

Bayesian-inference based recommendation in online social networks	In this content ratings are shared with friends. Conditional probability is used for calculating rating similarity.	1. Higher accuracy via friends' recommendation; 2. Solve the problem of large size of particle in collaborative filtering recommendation	There is a Cold start and rating sparseness problem.	Recommendation on online social networks
AWSR: Active Web Service Recommendation Based on Usage History	Web usage history and QoS are the main criteria for recommendation . Using this approach top k services are generated for users.	1.Higher recall ratio and accuracy; 2. Show the strength of the relationship between users.	Passive users reviews about the website is not considered. Usage history count is only used for ranking.	Website recommendation
Recommender System for Sport Videos Based on User Audiovisual Consumption	The recommendation is based on audiovisual consumption and does not depend on the number of users, running only on the client side.	This avoids the concurrence, computation and privacy problems of central server approaches in scenarios with a large number of users	Specific video fragment can not be recommended using this approach.	Sports recommendation
probabilistic personalized travel recommendation	For mining demographics for travel landmarks and paths people attributes and photos are used which are effective, and thus benefiting personalized travel recommendation services.	Only few parameters are used for similarity calculation.	Need to expand research work to include more attributes for accuracy and efficiency	Travel recommendation
Quality of service ranking prediction for cloud services	Rating based approaches and ranking based approaches are studied in this paper	users can obtain QoS ranking prediction as well as detailed QoS value prediction.	Applications in other field need further verification.	Quality of Service recommendation.

IV. PROPOSED SYSTEM ARCHITECTURE

General concept is given as follows. In figure 1. Proposed System Architecture overall concept of paper is established. Keywords are used to indicate both users preferences and candidate service quality. Similar users are then sorted using user based collaborative filtering algorithm. These similar users positive. negative reviews and sentiments in the text are differentiated. Sentiment analysis is used for score calculation. Top scoring services will be recommended first. Thus this Rank Boosting Approach recommends personalized ratings list to each user. It provides most appropriate top k ranking services to the user. Moreover, to increase scalability and efficiency MapReduce framework on Hadoop is used. General concept is as follow :



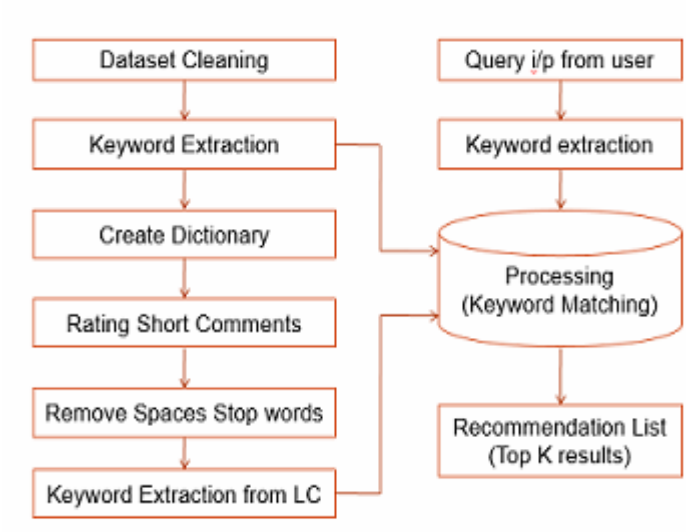


Figure 1. Proposed System Architecture

Methods that we are going to use are as follow:

1. Dataset Cleaning: In this method all keywords are extracted. First removal of stop words is done, then remove spaces. Replace lingo words with the root words. Then we are left with only keywords. Keywords are displayed as city name, hotel name, keywords. These keywords are stored in a database. Then keywords are stored in a file and retrieved using HDFS.

2. Creating Dependencies: Now we have a list of Keywords with us. For rating those words we have to create rating dictionary. This dictionary includes keyword and rate given to that keyword. When user enter any keyword, exact matching keyword is first searched then corresponding rating that we have given is obtained. For stop words stop.txt named dictionary is created. For checking lingo words lingo.txt is created.

3. Rating Short Comments: Short comments are retrieved and stored in a file. Each short comment is taken and then split by space is done. Each keyword is matched with the keyword stored in the rating.txt file. Calculation of all the rating values is done using Sentiment Analysis. Highest scoring hotel is ranked one and recommended first.

4. Keyword Extraction from Long Comments: Long comments include stop words, spaces , words in “ing” form so we have to remove all these things. To obtain keyword in root form Porter Stemmer algorithm is used. Stop words and spaces are removed using our own programming logic. Term Frequency (TF) is calculated. In cases where same keyword is extracted many times, reduce it to one. So more efficiency is obtained.

5. Recommendation : Keywords extracted from short comments, long comments and users preferences are stored in a dictionary. Rate all keywords. Calculate the overall score and then rank the hotel. Finally provide Recommendation list to users.

V. Implementation Results

Each time you have to write following commands and run Hadoop on terminal :

1. su – hduser :This command switches normal user to hduser.
2. hdfs namenode –format: This command is used to format namenode. Namenode contains metadata, i.e. it contains address of data which is present on the datanode. Each time you have to format the namenode so as new data loaded will be a fresh data and metadata storage is efficient. Each time a new id is provided to the namenode.
3. start-all.sh : This command is used to start working of all the daemons .
4. jps : It is java virtual machine process status tool. It is used to see the number of daemons running on your local machine.
5. stop-all.sh :This command is used to stop the working of all hadoop daemons.

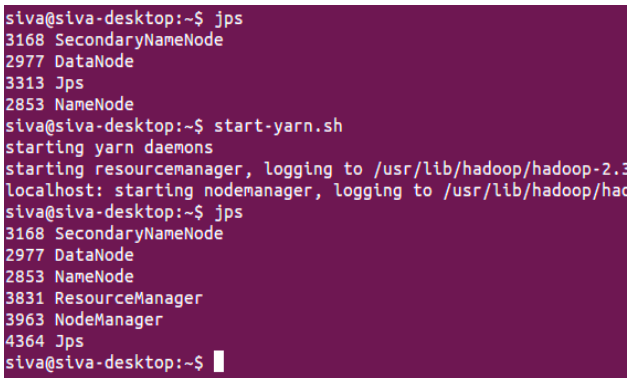


Figure 2. Run Hadoop on Terminal

Once you select the city from the combo box and enter keywords of your interest. These keywords are matched with the words stored in the database. The corresponding hotel reviews are passed to sentiment analysis. A total score is calculated using rating dictionary.

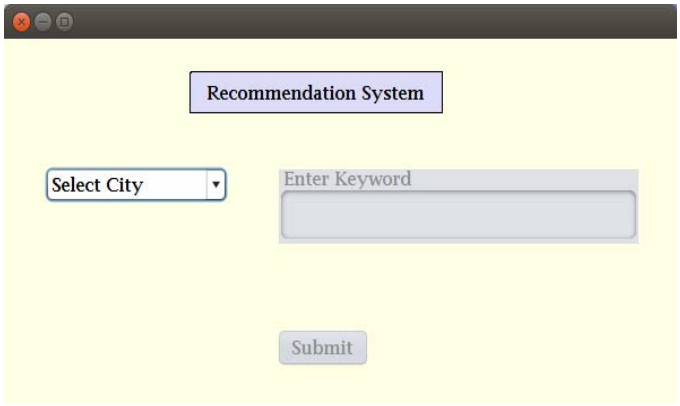


Figure 3. Recommendation System Home Page

We can add keywords of our own choice. These keywords entered if not in the lowercase they are converted to lowercase then compared with the keywords present in the database. Where a match is found that hotels sentiment score is calculated. If sentiment score is positive then it is recommended.

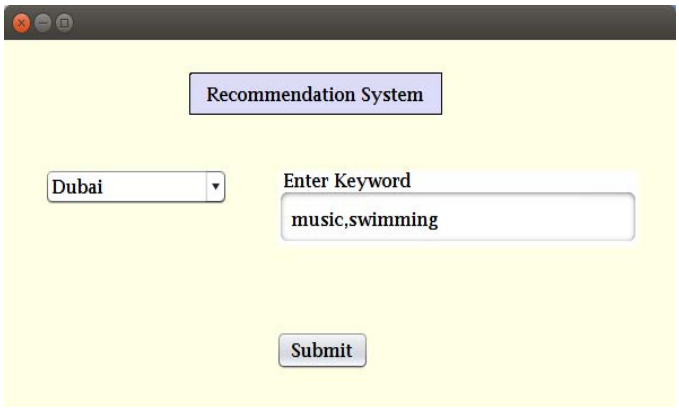


Figure 4. Run Application

Figure 5. Top-k Recommendations

VI. IMPLEMENTATION ANALYSIS

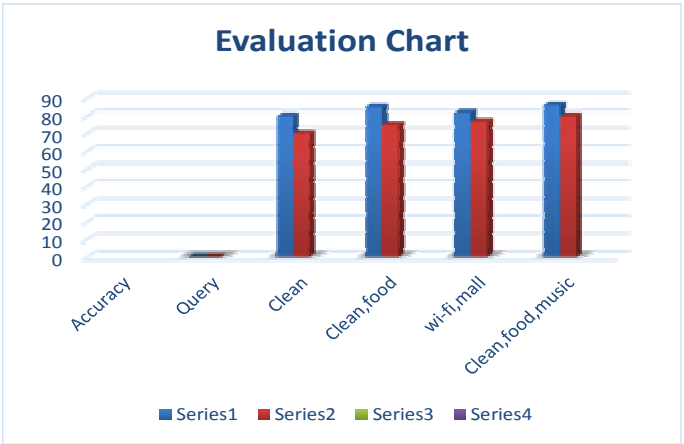


Figure 6. Implementation Analysis

Vertical axis represents the accuracy as compared with the traditional recommender systems. Horizontal axis represents the query inputted by the end user. Dataset of hotels is taken such as hotels in Chicago, New-Delhi, Beijing, London, New-York etc. In this, we have taken keywords such as “clean”, “clean, food”, “wi-fi, mall”, “clean,food,music”. The results obtained are more accurate than the traditional recommender system.

I. CONCLUSION

Recommendation is more dynamic and user friendly. More than two keywords can be used. Sentiment analysis is used to improve efficiency. As rate is calculated for all the positive long comments and recommendation is done. Hadoop is used to increase scalability and data security. Dataset can be used dynamically and accurate results are obtained.

REFERENCES

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen,” KASR:A Keyword-Aware Service Recommendation Method on Map Reduce for Big Data Applications” IEEE Transactions On Parallel And Distributed Systems, TPDS-2013-12-1141.

[2] X. Yang, Y. Guo, Y. Liu, “Bayesian-inference based recommendation in online social networks,” IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 4, pp. 642-651, 2013.

[3] G. Kang, J. Liu, M. Tang, X. Liu and B. cao, “AWSR: Active Web Service Recommendation Based on Usage History,” 2012 IEEE 19th International Conference on Web Services (ICWS), pp. 186-193, 2012.

[4] Yan-Ying Chen, An-Jung Cheng, “Travel Recommendation by Mining People Attributes and Travel Group Types From Community-Contributed Photos” IEEE Transactions on Multimedia, Vol. 15, No. 6, October 2013.

[5] M. Alduan, F. Alvarez, J. Menendez, and O. Baez, “Recommender System for Sport Videos Based on User Audiovisual Consumption,” IEEE Transactions on Multimedia, Vol. 14, No.6, pp. 1546-1557, 2013.

[6] Zibin Zheng, Xinmiao Wu, Yilei Zhang,Michael R. Lyu, Fellow,and Jianmin Wang,” QoS Ranking Prediction for Cloud Services” IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 6, June 2013.

[7] 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.

[8] Fuzhi Zhang, Huilin Liu, Jinbo Chao, “A Two-stage Recommendation Algorithm Based on K-means Clustering In Mobile E-commerce”, Journal of Computational Information Systems, Vol. 6, Issue 10, pp. 3327-3334, 2010.

[9] Brian McFee, Luke Barrington and Gert Lanckriet, “Learning Content Similarity for Music Recommendation” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 8, 2012.

[10] 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.

[11] D. Agrawal, S. Das, and A. El Abbadi, “Big Data and Cloud Computing: New Wine or Just New Bottles?” Proc. VLDB Endowment, vol. 3, no. 1, pp. 1647-1648, 2010.

[12] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” Comm. ACM , vol. 51, no. 1, pp. 107-113, 2005.

[13] S. Ghemawat, H. Gobioff, and S. T. Leung, “The Google File System,” Proc. 19th ACM Symp. Operating Systems Principles , pp. 29- 43, 2003

[14] Z. Luo, Y. Li, and J. Yin, "Location: A Feature for Service Selection in the Era of Big Data," Proc. IEEE 20th Int'l Conf. Web Service, pp. 515-522, 2013.

[15] B. Issac and W.J. Jap, "Implementing Spam Detection Using Bayesian and Porter Stemmer Keyword Stripping Approaches,"Proc. IEEE.