Sentiment Analysis Based Feedback Analysed Service Recommendation method For Big Data Applications

Khushboo R. Shrote, Mtech in CS and Engineering , GCOE Maharashtra,India Khushbooshrote166@gmail.com

Prof. A.V. Deorankar, Associate Professor in CS and Eng. GCOE Maharashtra,India avdeorankar@gmail.com

***Abstract*—**There is a rapid growth of customers, online information and services in recent years. Thus a personalized suggestion is required. Many traditional recommender systems are available but they lack in scalability, data security and efficiency. Recommender systems are the systems which has power to analyse historical data, reviews and can give individualized suggestions accordingly. Analysis of such a large amount of data that is "Big Data" is a major problem. Big Data is the term used when amount of data exceeds the current technology, method and theory to capture and process the data in required time. This big data is in the form of media logs, twitter and facebook logs. Thus no schema is defined, and it can not be analyzed using SQL. Cloud computing technology such as Hadoop is used to tackle this problem. Hadoop can analyse such data. There is no restriction on type of data it may be structured, unstructured or semistuctured. In this paper Feedback Analysis is done using Sentiment Analysis to Recommend services. Keywords are used to indicate what the users prefer, more than two keywords can be used. Hadooop can analyse data more efficiently and data security is also provided.

***Index Terms*—** recommender systems, Hadoop, Map Reduce, keywords, preferences, Sentiment Analysis, Reviews, Big Data.

1. INTRODUCTION

A. Background

Big Data analysis is one of the most important task to provide suitable services to customers. In most of the web sites the rating list and the recommendation list provided is almost same. Customers personal interest and requirements are not considered. In recent years services available, online data logs and customers are increasing exponentially. This data in the form of media logs, xml files, text files and twitter or facebook logs can not be analysed using SQL. In RDBMS using SQL we can fire query to obtain suitable data. This data is in the form of rows and columns called structured data. Our data is unstructured so SQL is of no use.

Hadoop is one of the cloud computing platform developed by Doug Cutting and team. Hadoop does not demand for structured data. Unstructured and semistructured data can also be analysed. Big data management is a challenge for IT industry. Hadoop provides solution to this problem as it is simple, robust, scalable.

B. Motivation :

In most existing service recommender systems for example hotel reservation systems the ratings given to services and the recommendation list provided to the customers are almost same. Customers personalized interest and requirements are not considered. Following is the example of Hotel Reservation system which illustrates the situation:

Example: Person A and person B are respectively browsing a Hotel website in Dubai. But the ratings given and the recommendation list provided to users are the same. Assume three hotels in Dubai. Hotel A, B and C. Hotel A has very good location. It is a good family hotel. Everything is situated nearby that is market, malls, airport etc. But the food quality disappoints. Hotel B has an excellent breakfast and other food. But market, malls and airport are situated far from hotel. Hotel C has very good transportation facilities. Food is good. Airport is nearby. According to ratings given B is better than C and A. A is better than C. But person A is interested in shopping and food he can adjust. According to ratings Hotel B will be the best one for him but according to his preferences Hotel A matches exactly his needs. So how to give individualized suggestions to users.

Motivated by above observations following method is proposed:

1. A feedback analyzed service recommendation method is proposed in this paper . This method is based on a user-based Collaborative Filtering algorithm.

2. In this method keywords extracted from query asked by the active user are used to indicate their preferences. These keywords are then matched with keywords extracted from reviews of passive users. Moreover, we implement it on Hadoop, which uses MapReduce as its computing framework.

3. Sentiment analysis is used to calculate score corresponding to reviews given. Highest value score is provided as the first in recommendation list.

Preliminary Knowledge

Recommendation System and Collaborative filtering

Recommender Systems are the systems which has capacity and power to give individualized suggestions by considering users different requirements. These are the subparts of information retrieval systems. Recommendation methods are classified into three categories as :

1.Content Based Recommendation
2.Collaborative Recommendation
3.Hybrid Recommendation

In content based recommendation items are recommended based on a comparison between the content of the items and a user profile on the website. In collaborative filtering similarity of users is computed. Similar users which buy the items is recommended. In hybrid recommendation two or more approaches are combined to gain more accuracy.

Hadoop and Map-Reduce :

Hadoop is a project developed by Doug Cutting and team that allows to store and process big data in a distributed manner across clusters of commodity hardwares. It is designed in such a way that it can easily scale up from single server to thousands of machines, which offers local computation and storage. Hadoop file system is developed as distributed file system design. Each block of data is of 64 KB or 128 KB or 256 KB or 1 MB. In normal file system 4 KB or 8 KB is allowed. So computation is much faster. Each block of data is replicated three times so data security is maintained. Replicated copies are updated time to time.

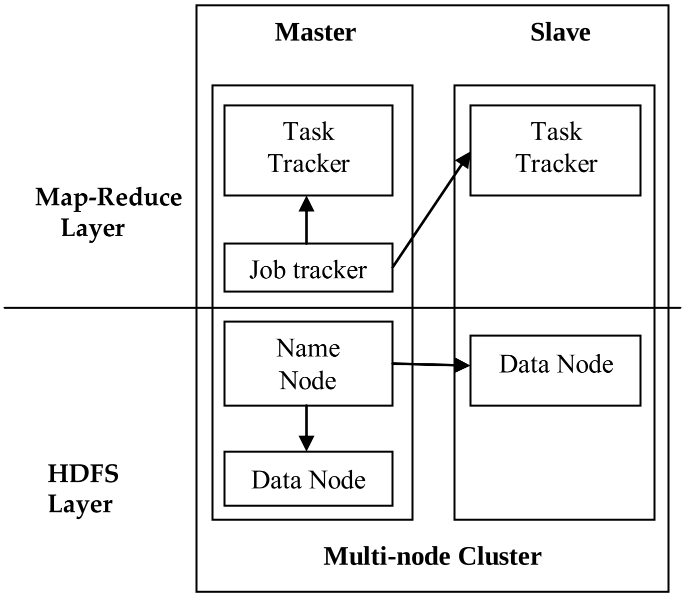

Fig 1. Hadoop Architecture

Hadoop is simple, robust, java compatible. Hadoop is cheap as only commodity hardwares are required. Commodity hardwares are the personal computers which are IBM compatible. Hadoop architecture is given in figure

Name Node: It is the master of the system. It maintains and manages the blocks which are present on the data nodes. Name node stores metadata.

Data Node: These are slaves which are deployed on each machine and provide the actual storage. Data node is responsible for serving read and write requests for the clients.

Job Tracker : It manages the jobs.

Task Tracker : It runs tasks that perform different parts of the job.

Map-Reduce architecture is inspired by Google. Map-Reduce refers to two distinct tasks that hadoop program perform. The first one is the Map job, which takes a set of data and convert it into another set of data, where individual elements are broken down into tuples (key/value pairs). Reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples.

2. LITERATURE REVIEW

Existing recommendation systems recommend services as per the ratings given. Users requirements and passive users reviews are not considered. Sentiments in the text have not been considered yet.

Shunmei Meng in 2014 proposed a KASR method for personalized recommendation. In this paper 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.They show that the proposed recommendation method is better than the existing traditional methods. Users positive and negative reviews are not differentiated and are not considered separately. Sentiments in the text is not considered for calculation.

Guosheng kang in 2012 proposed a paper on active web service recommendation. Web usage history and QoS are the main criteria for recommendation . Using this approach top k services are generated for users. Passive users reviews about the website is not considered. Usage history count is only used for ranking.

Xiwang yang in 2013 proposed a paper on Bayesian inference based recommendation in online social networks. In this content ratings are shared with friends. Conditional probability is used for calculating rating similarity. Based on similarity score ranking is done. They show that the proposed Bayesian-inference-based recommendation is better than the existing trust based recommendation. There is a Cold start and rating sparseness problem.

Faustuno Sanchez in 2012 proposed a paper on recommender system for sport videos, transmitted over the Internet and/broadcast, in the context of large-scale events, which has been tested for Olympic Games. The recommendation is based on audiovisual consumption and not 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, such as the Olympic Games. Whole video have to recommend. Specific video fragment can't be recommended using this approach.

Yan Ying Chen in 2013 proposed a paper on probabilistic personalized travel recommendation model. 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 reaserch work to include more attributes for accuracy

and efficiency.

Zibin Zheng in 2013 proposed a paper on quality of service ranking prediction for cloud services. Rating based approaches and ranking based approaches are studied in this paper, so that the users can obtain QoS ranking prediction as well as detailed QoS value prediction.

3. SYSTEM ARCHITECTURE

A. Overall Architecture of System :

In figure 2. general concept is given. In figure 3. 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 Map-Reduce framework on Hadoop is used. General concept is as follow :

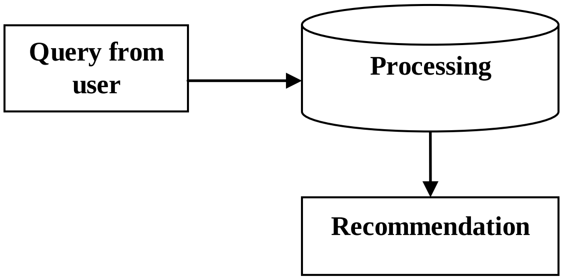

Fig 2. General Concept of Recommendation Method

Processing includes following steps :

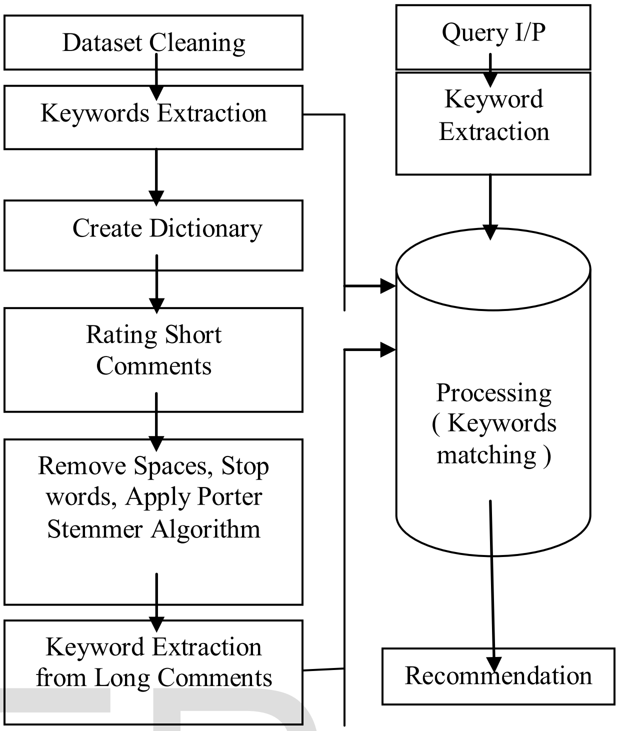
Dataset Cleaning
Keywords Extraction
Create Dictionary
Rating Short Comments
Remove Spaces, Stop words, Apply Porter Stemmer Algorithm
Keyword Extraction from Long Comments
Query I/P
Keyword Extraction
Processing (Keywords matching)
Recommendation

Fig 3. Overall System Architecture

The dataset of Hotels that we are going to refer contains short comments about hotels and long comments. First we will do analysis using short comments and provide recommendation. Then we will use long comments as input, do processing and extract keywords from comments. Keyword matching is done. Then rating according to keywords is done. Hotel which has highest rating is then recommended as top most rank choice.

B. Feedback Analyzed Service Recommendation Method

The main steps of FASR method are as follows :

1. Dataset Cleaning : In this method all keywords are extracted. First removal of stop words is done, then remove spaces. Then we are left with only keywords.
2. Creating Dictionary : 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.

3. Rating Short Comments : 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.

B.1. Dataset Cleaning :

In this method, keywords extracted from reviews must be stored somewhere. This storage place must be dynamic. That is number of keywords are not fixed this value increases always or sometime may decrease. So, we will use ArrayList class. Array have a fixed size. But ArrayList is flexible. It gets adjusted with respect to number of keywords. Then we will remove stopwords and spaces. Remaining word will be our keyword. Store this keyword in the array.Now we have a cleaned dataset.

B.2. Creating Dictionary :

A dictionary which contains all keywords along with their corresponding value is stored. This value between 1 to -1 is replaced with the keyword. Total Score is calculated using sentiment analysis.

Great	0.8
Brilliant	0.9
Recommendable	0.9
Highly	0.9
Attentive	0.8
Friendly	0.7
Smiling	0.6
Preposterous	-0.6
Mediocre	-0.4
Spacious	0.6
Free	0.4
Worst	-0.9
So-so	-0.2

Fig 4. Rating Dictionary

4. CONCLUSION

In this paper a feedback analyzed service recommendation method is proposed to recommend sevices to users. User based collaborative filtering algorithm is used to generate appropriate recommendations. Users can give more than one keyword as a preference. We have a huge dataset of hotels in the metro cities such as Dubai, London, Paris etc. First dataset cleaning is done. Stop words, spaces are removed then keywords are obtained. Exact matching keywords are found out from the dataset. We have Form the rating dictionary and have given rating values from -1 to +1. Sentiment Analysis is used for calculation. Hotel with highest rating value is ranked one and recommended first. This ranking is changeable. So we have to make updations in the rating dictionary as passive users reviews changes. So, Recommendation is dynamic and more realistic.

We are using Map-Reduce in java to reduce number of same keywords into one in the long. Finally we will run this project on Hadoop. Hadoop is an open-source framework designed by Doug Cutting and his team. Hadoop allows to store and process big data in a distributed manner across clusters of computers using Map-Reduce. It is designed to scale up from single servers to thousands of commodity machines, each offering local computation and storage.

Future Scope :

In future, research can be done in the area where a term appears other than the domain thesaurus.

REFERENCES

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen," KASR: A Keyword-Aware Service Recommendation Method on MapReduce 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 Internation-al Workshop on Knowledge Discovery and Data Mining, pp. 478-481, 2010.

[11] D. Agrawal, S. Das, and A. El Abbadi, "Big Data and Cloud Com- puting: 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 Region 10 Conf. (TENCON '09), pp. 1-5, 2009.

.