Hotel Recommendation System Using Hadoop Framework

KEYWORDS: Rating; Big data; Reviews; Hadoop; Map Reduce; Sentiment Analysis

Khushboo Ramesh Shrote
MTech student at Computer Science and Engineering Department Government College of Engineering Amravati Amravati, India

Anil V. Deorankar
Associate Professor at Computer Science and Engineering Deprtment Government College of Engineering Amravati, Amravati, India

ABSTRACT *Traditional recommender systems lack to give personalized recommendation to the end user. They lack in providing scalability and efficiency. Rating list and recommendation provided was almost same. So, in this paper a hotel recommendation system using hadoop framework is proposed. Hadoop mainly works in the area where big data appears. This big data is hard to capture and analyze. A review based service recommendation method is proposed to tackle this problem. This method is based on a user based collaborative filtering algorithm. Users having similar tastes are captured with the help of keywords they enter. Then Sentiment Analysis is applied on passive users reviews and a score is calculated. Top-k services are recommended to the end user. Experimental analysis shows that this method works more efficiently than traditional available methods.*

Introduction

After revolutionary success of web 2.0 information generated exponentially. The reason is it is continuously generated by various sources such as social media, power grid systems, stock exchange data, black box data, sensors, CCTV cameras. This data is unstructured or semi structured form. For processing of this large amount of data we need certain software tool. RDBMS is of no use as data must be structured and schema defined. Thus, Hadoop is worth to tackle this situation. As it can handle any type of data. Hive/HBase is the self database of hadoop which is more efficient in data retrieval than traditional software's.

BigData :

"It refers to 'Data' whose size is beyond the ability of current technology to process, handle and capture the data within particular instance of time". Big Data likewise conveys new opportunities and distinguishing troubles to industry and the academia, like most Big Data applications, the Big Data tendency likewise postures overwhelming effects on service recommender techniques. With the developing number of options for services, effectively recommending services that users favoured have turn into an imperative research issue. Service recommender framework has been indicated as important tools to help users manage services over-burden and give proper recommendations to them. As data is so huge, it is very difficult to manage a data and it also takes a much more time to generate results. So this is a drawback of using Big Data. But this issue can be dealt with the help of Hadoop, by using Hadoop a huge data can be analyzed in few seconds so Hadoop reduces the time.

Sentiment Analysis :

It is also known as opinion mining. It is used to check the positivity or negativity about any product or person. It can check which brand is the most famous. Also whether a person is seen positively or negatively on the web forum.

Hadoop :

Hadoop is a distributed computing framework and released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and also it can be efficient, reliable, scalable way to process data. Its core idea is to build on a large amount of cheap & efficient cluster hardware devices, in the form of software processing to pave the way of storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems. While referred to HDFS distributed fie systems, to improve fault tolerance in the form of software. When we compared with the traditional file system, it has low cost, easy to expand, and high fault tolerance features. Map Reduce is provided by Hadoop parallel computing model for handling large amounts of data calculation. Hadoop is scalable, it can easily meet the requirements of large-scale data need to handle on the PBLevel. In addition, the use of relatively low cost Hadoop cluster nodes require low internal computer, on any inexpensive computer can deploy In order to combined Hadoop with practical application better, you can also use some subprojects on the basis of Hadoop, such as Map Reduce technology, HBase distributed data storage systems, scalable data warehouse Hive, high-level data flow language Pig, high-performance distributed collaborative services Zoo Keeper and other major use Map Reduce technology and HBase data storage system to solve the problem of information retrieval. Hadoop is a cluster computing system which is data intensive. In this system incoming jobs are developed using the MapReduce programming model.

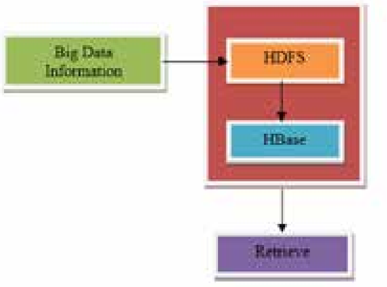


Figure 1. hadoop Framework
problem definition

In traditional recommender systems we might face the problems of scalability and efficiency in case of large scale data. Capturing and analyzing this data is hard. Traditionally personalized requirements of end users are not considered. The rating list itself is provided as the recommendation. Passive users review about any product or system is ignored. In this paper a Review Based Service Recommendation method is proposed, to achieve an efficiency of a Recommender System. It aims at presenting a personalized service recommendation list and recommending the most appropriate services to the users effectively.

literature survey

Recommender systems are developing as a powerful tool in both industry and academia. In [1] authors propose 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. Merits are 1. Scalable 2. More efficient than traditional methods. Demerits are Jaccard Coefficient method is not so accurate. Users positive and negative reviews are not differentiated. Sentiments in the text is not considered for calculation. In [2] authors propose an 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. Usage history count is only used for ranking. Merits are 1.Higher recall ratio and accuracy; 2. Show the strength of the relationship between users. Demerits are Passive users reviews about the website is not considered. Usage history count is only used for ranking. In [3] authors propose a 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. Merits are 1. Higher accuracy via friends' recommendation; 2. Solve the problem of large size of particle in collaborative filtering recommendation Demerits are 1. There is a Cold start and rating sparseness problem. In [4] authors propose 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. Merits are 1. 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. Demerits are 1. Specific video fragment can't be recommended using this approach. In [5] authors propose a 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. In [6] authors propose quality of service ranking prediction for cloud services. Rating based approaches and ranking based approaches are studied in this paper. Merits are 1. the users can obtain QoS ranking prediction as well as detailed QoS value prediction. Demerits are 1. Applications in other field need further verification.

System architecture

System Architecture of proposed method is shown in Figure 2. A dataset of hotels is taken for evaluation. It consists of date followed with short comment and a long comment.

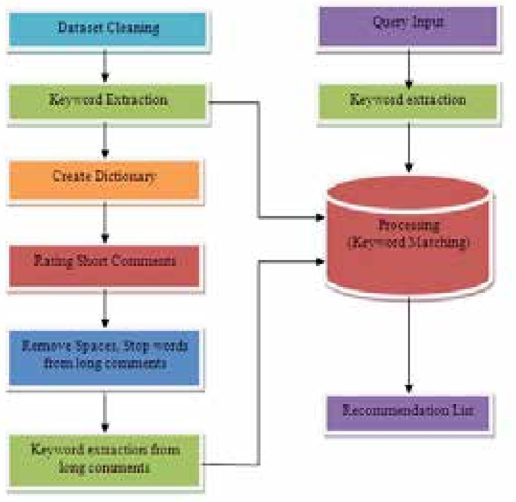
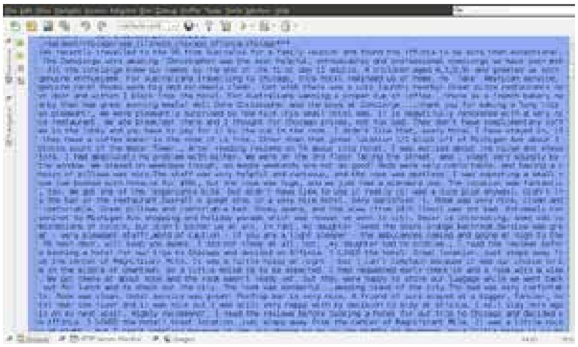


Figure 2. Proposed System Architecture

1. Preprocessing :

Ignore data upto first tab as it is a date. Remove stop words in the reviews to avoid affecting the quality of the keyword extraction in the next stage. And the Porter Stemmer algorithm is used to remove the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.



2. Keyword Extraction :

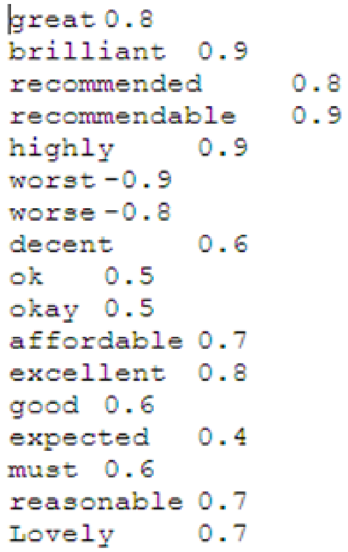
Each review will be transformed into corresponding keyword set. For example if a user type word transport the corresponding similar meaning words such as transportation, ship, move, convey must be present in the domain thesaurus.

For keyword in case of long comments only those long comments are preferred whose short comment rating is above the threshold value. Map Reduce algorithm can be applied so that same keywords must not be repeated.

Volume : 6 | Issue : 4 | April 2016 | ISSN - 2249-555X | IF : 3.919 | IC Value : 74.50

3. Create Dictionary :

Dictionary is a repository where a collection of keywords and their corresponding rating value is stored. Keywords are extracted from short and long comments. Their rating value is calculated by manual testing and it is set spaced with a tab. When we extract any keyword we pass it to check its rating value. This value is then obtained to use in score calculation in case of Sentiment Analysis.



```
great 0.8
brilliant  0.9
recommended      0.8
recommendable    0.9
highly     0.9
worst -0.9
worse -0.8
decent     0.6
ok    0.5
okay  0.5
affordable 0.7
excellent  0.8
good  0.6
expected   0.4
must  0.6
reasonable 0.7
Lovely     0.7
```

4. Rating Short Comments :

All short comments of a particular hotel is captured and stored in a file. Name that file as file 1. Then keywords are extracted their corresponding rating values are checked. Sentiment Analysis is applied and a total rate value is calculated. If this value is above threshold value then that hotel is included otherwise ignored.

5Recommendation List :

After Keyword matching recommended hotels are generated. Passive reviews of each hotel are considered Sentiment analysis is applied to these reviews and a score is calculated. If score is above threshold value then it is granted for further evaluation otherwise rejected. According to total sentiment score top-k services are recommended as a list to the end users.l.

6. conclusions

In this paper a review based service recommendation method is proposed to recommend services 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 formed 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 user's 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.

7. 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.