Recommendation System: A Big Data Application

Article · September 2016

CITATIONS
4

READS
2,496

2 authors, including:

K R Remesh Babu Raman
Government Engineering College, Sreekrishnapuram Palakkad
46 PUBLICATIONS **239** CITATIONS

Open access Journal International Journal of Emerging Trends in Science and Technology
Impact Factor: 2.838 INC-BEAT 2016

Recommendation System: A Big Data Application

Reshma M[1], K.R Remesh Babu[2]

1(Information Technology ,GEC Idukki,India)

2(Information Technology, GEC Idukki,India)

ABSTARCT

Recommender systems have been shown as tools for providing appropriate recommendations to users. The amount of customers, services and online information is growing up day by day,as a result yielding the big data analysis becomes a problem for service recommender systems. Consequently, existing recommender systems suffer from scalability and inefficiency problems while processing or analysing such large-scale data. Moreover, the same ratings and rankings systems used in most of traditional recommender systems in which without considering diverse users' preferences, as a result they fails to meet users' personalized requirements. In this paper, we propose a user aware method to address the above challenges. It aims at presenting a personalized listof recommendationserviceand recommending the most appropriate services to the users effectively. Specifically, keywords indicate the preferences of corresponding user's,in order to improve its scalability and efficiency in big data environment.Itis implemented on Hadoop, a widely adopted distributed computing platform using the MapReduce parallel processing paradigm. Finally, Future experiments are conducted on real-world data sets and configuring the multimode. The results demonstrate that it significantly improves the accuracy and scalability of service recommender systems over existing approaches

Recommender system,: Big Data, Hadoop, Keyword , MapReduce, Preference

1 Introduction

The technology is grown drastically. New technologies, devices, and communication means like social networking sites, the amount of data produced by human is growing day by day. Big data means really a big data, it is a collection of large datasets that cannot be processed using our old computing techniques. Big data is not merely a data, rather it has become a subject, which involves various tools, techniques and frameworks. Big data is really critical to our life. Its emerging as one of the most excellent technologies in current world. The concept big data having some benefits, which are familiar to us. Using the information in the social media like preferences and product perception of their consumers, The companies are planning their goal based on the store .In hospital using previous medical history of patients, for providing better service. Massively Parallel Processing database systems and MapReduce that provide analytical capabilities for complex analysis that may touch most or all of the data. Accuracy in

big data may lead to more confident decision making, and better decisions can result in good operational efficiency, cost reduction and reduced risk. Here using the recommendation System by trip advisor data set. Based on user similarity recommending the best hotel.

2. Scope And Objectives

A recommender system would calculate the similarities between two users namely acive and previous. Personalized rating will calculating based on the similarity, By this generating the best k preferences. Recommender systems trying to reduce information overload by this retain customers by sending appropriate recommended items from a universal set based on user preferences. A preference indicates an individual mental state concerning a subset of items from the universe of alternatives. Individuals having preferences based on their experience with the relevant items. Recommender systems calculate the conditional probability of the items in a probabilistic space. The similarity between user's are calculate based on concept of vector space modal. Recommendation seekers, computer programmers, and recommendation provider's recommender Systems in current use Amazon.

3. Literature Survey

In [1], the authors propose a Bayesian inference based recommendation system for online social networks. They show that the proposed Bayesian inference based recommendation is better than the existing trust based recommendations and is comparable to Collaborative Filtering recommendation. In ,Adomavicius and Tuzhilin give an overview of the field of recommender systems and describe the current generation of recommendation methods. They also describe various limitations of current service recommendation methods, and discuss possible extensions that can improve recommendation capabilities and make recommender systems applicable to an even broader range of applications. Most existing service recommender systems are only based on a single numerical rating to represent a service's utility as a whole. In fact, evaluating a service through multiple criteria and taking into account of user feedback can help to make more effective recommendations for the users. The authors of implement a CF algorithm on Hadoop. They solve the scalability problem by dividing dataset. But their method doesn't have favourable scalability and efficiency if the amount of data grows. Presents a parallel user profiling approach based on folksonomy information and implements a scalable recommender system by using Map Reduce and Cascading techniques. propose a large scale video recommendation system based on an item-based CF algorithm. They implement their proposed approach in Qizmt, which is a .Net Map Reduce framework, thus their system can work for large scale video sites. Generally speaking, comparing with existing methods, KASR utilizes reviews of previous users to get both of user preferences and the quality of multiple criteria of candidate services, which makes recommendations more accurate. In Hybrid Recommender Systems.Survey and

Experiments [2] represent user preferences for the purpose of suggesting items to purchase or examine. They have become fundamental applications in electronic commerce and information access, providing suggestions that effectively prune large information spaces so that users are directed toward those items that best meet their needs and preferences. A variety of techniques have been proposed for performing recommendation, including content-based, collaborative, knowledge-based and other techniques. To improve performance, these methods have sometimes been combined in hybrid recommenders. This paper surveys the landscape of actual and possible hybrid recommenders. Further, we show that semantic ratings obtained from the knowledge-based part of the system enhance the effectiveness of collaborative filtering. In the paper New Recommendation Techniques for Multi Criteria Rating Systems [3] traditional single rating recommender systems have been successful in a number of personalization applications, the research area of multi criteria recommender systems has been largely untouched. In order to take full advantage of multi criteria ratings in various applications, new recommendation techniques are required. In this paper we propose two new approaches the similarity based approach and the aggregation function-based approach to incorporating and leveraging multi criteria rating information in recommender systems. We also discuss multiple variations of each proposed approach, and perform empirical analysis of these approaches using a real world dataset.

Table 1 Merits and demerits of existing systems [4],[5]

Author	Advantage	Disadvantage
X.Yang,Y.Guo et al	Share rating with friends More accurate than traditional approach	Rating similarity between pair of friends is measured by a set conditional probability derived from mutual rating Muticriteria not supported
G.Adomaqvicius et al	Support Multicriteria rating	Not supported for broader range of application
Y.Kown eta al	Multicriteria rating Effective	Not scalable
Z.D. Zhao et al	Implemented on Hadoop Solve scalability problem	Does not have favorable scalability and efficiency if the amount of data grown

4. System Design

The Data set was downloaded from UCI repository. Trip advisor data set was using. The attributes of this dataset are Hotel name, Address, Date of the Review and Users review. Then the Active user gives their preferred item by keywords. That item was compared to the previous user preference. Then generate the Recommendation to the Active user. Then similarity checking, Ranking and Rating, Recommendation Generation are the major steps. The system having keyword candidate list.Hear it crated manually.The problem implemented by] using the concept of MapReduce [6].

4.1 Preferences of an active user and previous users

We having mainly two users active and previous user. Active user can need recommendation and previous user who already in the data set. An active user can give preferences about the services by selecting keywords from a list of keywords, which indicate his.

The preferences of a previous user are extracted from his/her reviews for the service according to the list of keywords and domain thesaurus.

a) Pre-process: HTML tags from hotel link and stop words in the reviews snippet collection should be removed. The Porter Stemmer algorithm for removing the morphological and in flexional endings from words in English.

b) Keyword extraction: Review will be transformed into a corresponding keyword set according to the list of keywords and domain thesaurus.

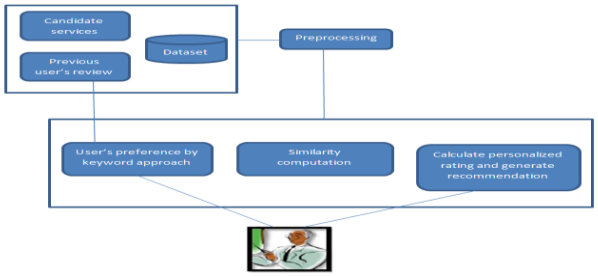

Fig 1. Main steps

Two similarity computation method are used, that are approximate and exact similarity computation. For approximate similarity method the weights of the keywords in the preference keyword set are not considering, while the exact similarity computation method will considering the weight of the keyword.

4.2 Approximate similarity computation

A best method for comparing the similarity and diversity of sample sets, Jaccard coefficient 1, is applied in the approximate similarity computation.

$$sim(APK, PPK) = \frac{|APK \cap PPK|}{|APK \cup PPK|} \qquad (1) \quad [7]$$

4.3 Exact similarity computation

Hear using Cosine based similarity computation. [8]. After checking the consistence of the matrix, then we calculate the weight

$$wi = \frac{1}{m}\sum_{j=1}^{m}\frac{aij}{\sum_{k=1}^{m} akj} \qquad (2) \quad [7]$$

Where *aij* is the relative importance between two keywords, m is the number of the keywords in the preference keyword set of the active user. The weight vector of the preference keyword set of a previous user can be decided by the term frequency/inverse document frequency (TF-IDF) measure. In the TF-IDF approach, to calculate the preference weight vector of a previous user u', "all reviews" by user u' should be collected. Here, "all reviews" contain the reviews by user u' for the candidate services and similar services not in the candidate services.

$$TF = \frac{Npki}{\sum_{g} Npki} \qquad (3) \quad [7]$$

where $Npki$ is the number of occurrences of the keyword pki in all the keyword sets of the reviews commented by the same user u', g is the number of the keywords in the preference keyword set of the user u'. The inverse document frequency (IDF) is obtained by dividing the number of all reviews by the number of reviews containing the keyword pki.

$$IDF = log\frac{|R'|}{|r' : pki \epsilon r'|} \qquad (4) \quad [7]$$

where $|R'|$ is the total number of the reviews commented by user u' , and $|r' : pki \epsilon r'|$ is the number of reviews where keyword pki appears. So the TF-IDF weight of the keyword pki in the preference keyword set of user u'can be decided by the following function:

$$wki = TF \times IDF \quad (5) \quad [7]$$

Algorithm 2 SIM-ESC (Exact Similarity Computation)

1: **Input:** The preference keyword set of the active user APK The preference keyword set of a previous user PPK_j
2: **Output:** The similarity of APK and PPK_j, $sim_{ESC}(APK, PPK_j)$
3: for each keyword k_i in the keyword-candidate list
4: if $k_i \in APK$ then
5: get $\overrightarrow{W}_{AP,i}$
6: else $\overrightarrow{W}_{pp_j,i} = 0$
7: end if
8: if $k_i \in PPK_j$
9: get $\overrightarrow{W}_P P_j, i$
10: else $\overrightarrow{W}_{PP_j,i} = 0$
11: end if
12: end for
13: get PPK_j, $sim_{ESC}(APK, UPK_j)$
14: return the similarity of APK and PPk_j ,$sim_{ESC}(APK, UPK_j)$

Fig 3.Exact similarity computation [7]

4.4 Calculate personalized ratings and generate recommendations:

Given a threshold γ if *sim(APK, PPK)* <γ , The thresholds given in two similarity computation methods are different, which are both empirical values. Once 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 with the highest ratingwill be recommended to him/her.

5. Experimental Evaluation

5.1 Experiment Setup and Datasets

Hear implanting a recommendation system. In which using ASC and ESC.ASC based on Jaccard coefficient and ESC giving recommendation based on cosine based similarity.ESC considering weight of the keyword based on TF-IDF values.Technically, our experiments are conducted in a Hadoop platform. And to evaluate the accuracy and scalability of our system, two kinds of dataset are adopted in the experiments a real dataset and a synthetic dataset. Most commonly a data set corresponds to the contents of a database table, or a single statistical data matrix, where every column of the table represents a particular variable. The data set may comprise data for one or more members, corresponding to the number of rows. TripAdvisor was an early adopter of user generated content. It provide most of the content, and the website is supported by an advertising business.The full TripAdvisor dataset consists of 235,793 hotel reviews collected over a period of one month. In addition to the review text, each review comes with a hotel identifier, an overall rating .We are using the following keywords Rooms, Cleanliness, Value, Service, Location, and Business.

5.2. Tools Useds

Hadoop is working on different flavours of GNU/Linux platform. Therefore, we want to install a Linux operating system for setting up Hadoop environment. Java is the main prerequisite for Hadoop. After the completion of installation configuring data node,Name node,Recourse manager, Secondary name node, Node manager. Hadoop accessing default port 50070.Create directory and upload the dataset. It having the size of 255.2MB and having 1759 items. This take more time for data pre-processing.

A dataset file mainly contains overall rating,average price,url,author,comment,date,overall,etc.From this data performing pre-processing the getting only our required information such as address, hotel name, author and comment and rating.

5.3 Experiments

5.3.1 Execution Time

Hear weanalysing ASC and ESC.The experiment is conducted on 2MBTripadvisor dataset.The preferences are 'value,food,room' for this experiment. The execution time ASC shown in the table 2 and ESC in 3

Table 2 Execution time for ASC

Dataset size	Time in sec
2 MB	10.4966
4 MB	11.5308
6 MB	12.5572
8 MB	14.7214
10 MB	18.1098

Table 3 Execution time for ESC

Dataset size	Time in sec
2 MB	10.6722
4 MB	12.5642
6 MB	13.9228
8 MB	16.1506
10 MB	20.3842

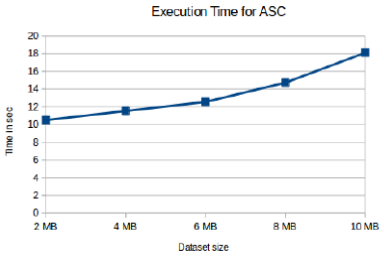

Fig 5.Execution time for ASC

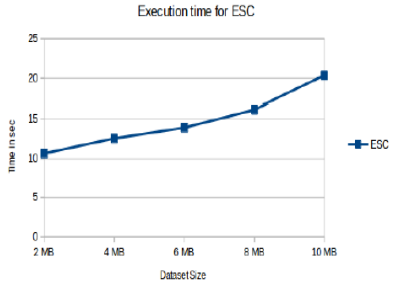

Fig 6. Execution time for ESC

Dataset size	Time for ASC	Time for ESC	Percentage change in time
2 MB	10.4966	10.6722	0.8295
4 MB	11.5308	12.5642	4.2888
6 MB	12.5572	13.9228	5.157
8 MB	14.7214	16.1506	4.6294
10 MB	18.1098	20.3842	5.9084

Table 4 Percentage change in time

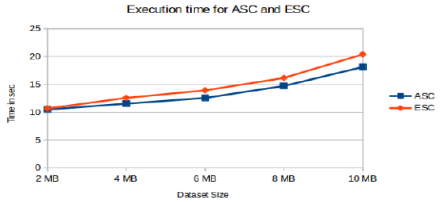

Fig 7.Combination of ASC ESC

5.3.1 Personalized rating

Both ASC and ESC are finding the personalized rating pr.Hear mainly 10 keywords are used.The program running based on 2 MB dataset.These 10 keyword having differed Pr .It depends on dataset and similarity

between active and previous user.

Table 5 Personalized rating

Keywords	Personalized rating for ASC	Personalized rating for ESC
Value	3.378	3.616
Transportation	3.67	3.83281
Cleanliness	3.599	3.7541
Food	3.6195	4.15
Room	2.876	3.8289
Service	3.2828	3.4495
Shopping	3.98	4.0132
Fitness	4.396	3.7848
Family	4.5568	4.20148
Environment	3.9924	4.008

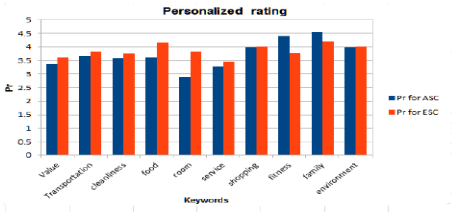

Fig 8.Personalized rating

5.4.TF and IDF

The weight vector of the preference keyword set of a previous user can be decided by the term frequency/inverse document frequency (TF-IDF) measure.

Table 6 TF-IDF values

Review	TF	IDF
1	1.0	0.6931
2	1.0	0.6931
3	1.0	0.6931
4	0.2857	0.6931
5	0.1428	1.3862
6	0.2857	1.0986
7	0.1428	0.6931
8	0.1428	0.6931
9	1.0	2.1972
10	0.2	0.6931

Table 7 Similarity, Rating, Overall rating(ASC)

Review	Similarity(avg)	Rating (avg)	Overall rating	Personalized rating
1	0.2606	2.2	3	2.26
2	0.2528	2.33	3	2.41
3	0.251	2.54	4	2.59
4	0.25	2.55	3.5	2.57
5	0.26	2.65	4	2.68

5.4.1. Similarity,Rating,Overall rating

Similarity calculated based on ASC and ESC.Similarity in ASC based on active and previous user preferences .ESC based on preference weight vector. The table 7 and 8 shows the analysis.When running the program 2 MB dataset is used.

Table 8 Similarity, Rating, Overall rating(ESC)

Review	Similarity(avg)	Rating (avg)	Overall rating	Personalized rating
1	0.44	2.5	3	2.53
2	0.43	2.5	3	2.53
3	0.519	5	4	3.27
4	0.538	3.29	3.5	3.2
5	0.338	3.2	4	3.2

6. Conclusions And Future Enhancemants

In this paper discussed about recommendation

system.Keywords are representing the preference of users.User based CF algorithm will be the back born.Specifically a list of keywords and domain thesaurus help to get preferences.This recommendation system considering active users he/she need the recommendation and passive user who already login in to the trip advisor site.Based on the similarity between theses users calculate personalized rating and finally generate best k recommendations.

The experiments are conducting on real world data set and including the concept of IoT. Sensors are locating the location and the recommendation are sending to the users mobile phone. Using any cloud provider (like AWS) the experiments are running in multimode hadoop. checking the accuracy with original trip advisor and collecting feedback from face book, twitter etc.Also considering sentimental analysis in pre processing stage.By this we get only positive reviews.

7. Refrences

[1] 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.

[2] R. Burke, "Hybrid Recommender Systems: Survey and Experi-ments," User Modelingand User-Adapted Interaction, Vol. 12, No.4, pp. 331-370, 2002.

[3] K. Lakiotaki, N. F. Matsatsinis, and A. Tsoukis,"Multi-Criteria User Modelingin Recommender Systems", IEEE Intelligent Systems, Vol. 26, No. 2, pp. 64-76, 2011.

[4] G. Adomavicius, and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the Stateof- the-Art and Possi-ble Extensions," IEEE Transactions on Knowledge and Data Engi-neering, Vol.17, No.6 pp. 734-749, 2005.

[5] G. Adomavicius, and Y. Kwon, "New Recommendation Techniques for MulticriteriaRating Systems," IEEE Intelligent Systems, Vol. 22, No. 3, pp. 48-55, 2007.

[6] Y. Jin, M. Hu, H. Singh, D. Rule, M. Berlyant, and Z. Xie, "MyS-pace Video Recommendation with Map-Reduce on Qizmt," Pro-ceedings of the 2010 IEEE Fourth International Conference on Se-mantic Computing, pp.126-133, 2010

[7] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, *Senior Member, IEEE "*

KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications"

[8] Qian, Gang. Sural, Shamik. Gu, Yuelong. Pramanik, Sakti. "Similarity between Euclidean and cosine angle distance for nearest neighbor queries", (2003).