Preference Aware Service Recommendation Using Collaborative Filtering Approach

Ms. Sneha.V
Student (M Tech) of Dept. of Computer science & Engineering
Sai Vidya Institute of Technology, Bangalore
Bangalore, 560064, India
snehayadav.sn@gmail.com

Mr. S Deepak Raj
Associate professor,
Dept. of Computer science & Engineering
Sai Vidya Institute of Technology, Bangalore,
Bangalore,560064 ,India
Deepak.raj@saividya.ac.in

Abstract- Service recommendations are shown as remarkable tools for providing recommendations to users in an appropriate way. In the last few years, the number of customers, online information and services has grown very rapidly, resulting in the big data analysis problem for service recommendation system. Consequently, there is scalability and inefficiency problems associated with the traditional service recommendation system which suffers in processing or analyzing large-scale data. Moreover, most of available service recommendation system gives the same rankings and ratings of services to different users without any considerations of many user's preferences, and hence it fails to reach user's personalized requirements. In this paper, we have proposed a Preference-Aware Service Recommendation method, to overcome the above challenges. It aims at recommending the most appropriate and preferred services to the users and provide a personalized service recommendation list in an effective way. Here, users' preferences are captured as keywords, and a user-based Collaborative filtering approach is adopted to create appropriate recommendations. A widely-adopted distributed computing platform, Hadoop is used for the implementation of this approach, which improves its efficiency and scalability in big data environment, using the MapReduce parallel processing method.

Keywords*-recommender system, preference, keyword, service, efficiency, scalability, collaborative filtering, big data, Map-Reduce, Hadoop.*

*****

I. INTRODUCTION

A. Big Data

Big data usually contains sets of data with sizes which are beyond the ability of software tools that are commonly used to capture, curate, manage, and process data within a tolerable time limit [1]. Big data is a set of technologies and techniques, requires different and new forms of integration to uncover large hidden values from huge datasets that are of a massive scale, complex and diverse.

The challenges with big data include capture, analysis, sharing, search, transfer, storage, privacy violations and visualization. Big data is difficult to work when most relational database management systems and desktop statistics are used.

B. Cloud Computing

Cloud computing focuses on computations over a scalable network of nodes and sharing the data. The important goal of cloud computing is to share the resources, such as platform, infrastructure, business process and software. There are many tools available for cloud computing, such as Mahout (http://mahout.apache.org/), Hadoop (http://hadoop.apache.org/), the Dynamo of Amazon.com, MapReduce of Google.

C. Hadoop

The Hadoop, a distributed computing platform is a batch processing system for a groups of nodes that provides the most Big Data activities of analytics because it groups two sets of functionality mostly needed to deal with unstructured large datasets like, Map-Reduce modeling and distributed file system (DFS) It is a project written in Java by the Apache Software Foundation to support data intensive distributed applications. Hadoop enables applications to work with peta bytes of data and thousands of nodes. The biggest contributor of Hadoop has been the search g engine Yahoo, where it is extensively used widely in the business platform.

The rest of the paper is organised as follows: section II explains about the proposed approach, section III about the implementation, section IV about the experimental evaluation and section V about the Conclusion.

II. PROPOSED APPROACH

A. Preference Aware service recommendation system

In this method, keywords are used to indicate the quality of candidate services and user's preferences. Two data structures, "keyword-candidate list" which is a set of keywords about multi-criteria of the candidate services and preferences of users[2]. And A domain thesaurus is created as keyword-candidate list's reference work that lists words that are grouped together as per the similarity of keyword.

Jaccard coefficient is a measurement of asymmetric information on binary or non-binary variables and used to find a similarity between the Previous user's preference keyword (PPK) and active user's preference keyword (APK). Exact similarity computation is calculated using cosine-based approach which is similar to the (VSM) Vector Space Model that is used for information retrieval[3]. The preference keyword sets of the previous users and active user will be transformed as weight vectors of *n*-dimensional respectively, namely as preference weight vector.

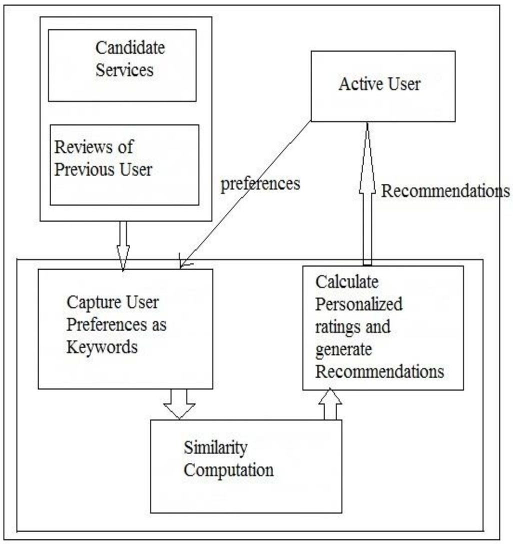


Fig 1. The Architecture of the proposed service recommendation system

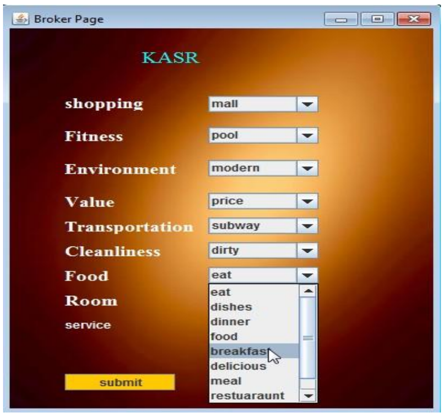
The preference keyword set's weight vector of a previous user is considered by *the term frequency/inverse document frequency* (*TF-IDF*) measure, one of the well-known measures in Information Retrieval for specifying the weight of the keywords.

In the last step, further filtering will be conducted based on the similarity of the previous users and active user. Once the most similar set of users are found, each candidate service's personalized ratings can be calculated for the active user. And finally, a personalized recommendation list of services will be presented to the active user and the services with the highest ratings will be recommended to him/her.

III. IMPLEMENTATION

To improve the scalability and efficiency, it is implemented on Hadoop platform in a MapReduce framework. Appropriate recommendations are generated using a user-based Collaborative filtering algorithm. A domain thesaurus is created which is a keyword-candidate list's reference work that lists the words grouped together according to the keyword meaning similarity.

In the first step, the preferences of previous users and active users are formalized into their corresponding preference keyword sets. An active user can give their candidate services preferences by choosing keywords that reflect the quality criteria of the services they are concerned about. The previous user's preferences for a candidate service are extracted from their reviews for the service according to the domain thesaurus and keyword-candidate list.


Broker Page

KASR

shopping mall
Fitness pool
Environment modern
Value price
Transportation subway
Cleanliness dirty
Food eat
Room
service

Fig 2. Service request page

In the next step, an exact similarity computation method and approximate similarity computation method are calculated for each keywords. In the approximate similarity computation, Jaccard coefficient method is used.

The preferences similarity between the previous user an active user based on Jaccard coefficient is calculated as follows:

$$sim(APK, PPK) = Jaccard(APK, PPK) = \frac{|APK \cap PPK|}{|APK \cup PPK|}$$

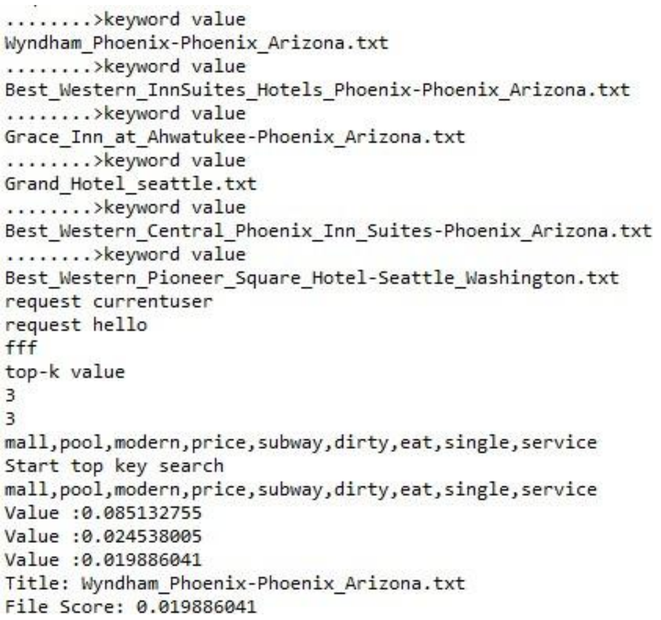
In the exact similarity computation, a method similar to the Vector Space Model calledcosine-based approach is applied that is used in the information retrieval process. The preference keyword sets of the previous users and active user are transformed into *n*-dimensional weight vectors respectively, as preference weight vector list. For each keywords in keyword candidate, if that keyword belongs to APK then a preference weight vector Wap is calculated for the keywords using:

$$w_i = \frac{1}{m} \sum_{j=1}^{m} \frac{a_{ij}}{\sum_{k=1}^{m} a_{kj}}$$

Where m is the total number of the keywords in the preference keyword set of the active user, aij is the relative importance between the two keywords, The preference keyword set's weight vector is shown by *the term frequency/inverse document frequency* (*TF-IDF*) by using formula:

$$w_{pk_i} = TF \times IDF = \frac{N_{pk_i}}{\sum_g N_{pk_i}} \times \log \frac{|R'|}{|r' : pk_i \in r'|}$$

Where, N is the number of keyword occurances *pki* in all the review's keyword sets commented by the same user *u'*, *g*-number of the keywords in the preference keyword set of the user *u'*. |*R'*|- number of the reviews commented by user *u'*, and |*r'*: *pki*∈ *r'*| - number of reviews where keyword *pki* appears.



```
........>keyword value
Wyndham_Phoenix-Phoenix_Arizona.txt
........>keyword value
Best_Western_InnSuites_Hotels_Phoenix-Phoenix_Arizona.txt
........>keyword value
Grace_Inn_at_Ahwatukee-Phoenix_Arizona.txt
........>keyword value
Grand_Hotel_seattle.txt
........>keyword value
Best_Western_Central_Phoenix_Inn_Suites-Phoenix_Arizona.txt
........>keyword value
Best_Western_Pioneer_Square_Hotel-Seattle_Washington.txt
request currentuser
request hello
fff
top-k value
3
3
mall,pool,modern,price,subway,dirty,eat,single,service
Start top key search
mall,pool,modern,price,subway,dirty,eat,single,service
Value :0.085132755
Value :0.024538005
Value :0.019886041
Title: Wyndham_Phoenix-Phoenix_Arizona.txt
File Score: 0.019886041
```

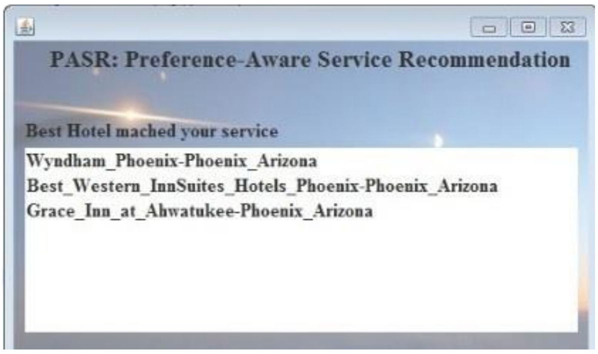
Figure3. Personalized rating calculation using similarity computations

Finally, the personalized ratings of each candidate service can be calculated once the set of most similar users are found. A average weighted approach for calculating *pr*, the personalized rating of a service for the active user is calculated using the formula:

$$pr = \bar{r} + k\sum_{PPK_j \in \hat{R}} sim(APK, PPK_j) \times (r_j - \bar{r})$$

$$k = 1 \Big/ \sum_{PPK_j \in \hat{R}} sim(APK, PPK_j)$$

rj is the rating of the corresponding *PPKj* review; R^ denotes the set of the remaining previous user's keyword sets, and r- is the average ratings of the candidate service.



PASR: Preference-Aware Service Recommendation

Best Hotel mached your service

Wyndham_Phoenix-Phoenix_Arizona
Best_Western_InnSuites_Hotels_Phoenix-Phoenix_Arizona
Grace_Inn_at_Ahwatukee-Phoenix_Arizona

Fig 4. Preference recommendation list

Then the services can be ranked by the personalized ratings and a personalized service recommendation list can be presented to him/her. Here we assume that the services with higher ratings are more preferable to the user. So the services having the highest ratings will be recommended to the active user.

IV. EXPERIMENTAL EVALUATION

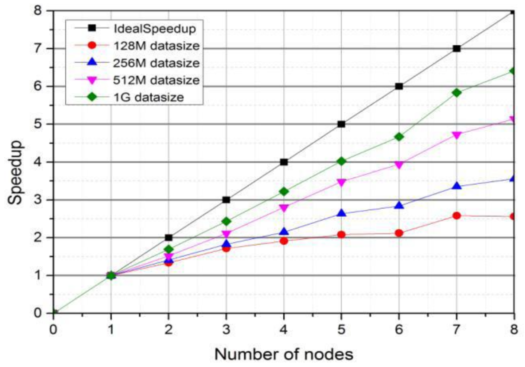


Fig 5. Experimental evaluation

To verify the scalability of PASR, experiment is conducted using cluster of nodes which are ranging from 1 to 8. There are 4 datasets used in this experiments(128M, 256M, 512M and 1G datasize). It is observed that the speedup of KASR increases linearly with the growth of the no of nodes. Meanwhile, the bigger dataset obtained a better speedup. When the number of nodes is 8 and the data size is 1G, the value of the speedup reaches 6.412, that is 80.15% (6.412/8=80.15%) of the ideal speedup. So, the experimental result shows that has good scalability over "Big Data" PASR on Map-Reduce in Hadoop platform and performs better with bigger dataset.

V. CONCLUSION

In this proposed system, keywords shows user preferences and collaborative filtering approach is used to provide appropriate recommendation lists of services to the user. For faster calculations and scalability, it is implemented on Hadoop Map-Reduce framework which gives better scalability and accuracy.

REFERENCES

[1] Manyika, M. Chui, B. Brown, et al, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

[2] K. Abhishek, S. Kulkarni, V. Kumar, N. Archana and P. Kumar, "A Review on Personalized Information Recommendation System Using Collaborative Filtering," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 2, no. 3, pp. 1272-1278, 2011.

[3] Heung-Nam Kim1, Ae-Ttie Ji1, Cheol Yeon1, and Geun-Sik Jo2,"] A User-Item Predictive Model for Collaborative Filtering Recommendation" 2008.

[4] Xuecong ZENG, Longsheng CAI and Tomohiro MURATA, "Preference based Recommendation Method by a hierarchy process and a Case Study of Smartphone", 2011.

[5] M. Bjelica, "Towards TV Recommender System Experiments with User Modeling," IEEE Transactions on Consumer Electronics, Vol. 56, No.3, pp. 1763-1769, 2010.