IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-3, Issue-4 E-ISSN: 2347-2693

A Survey on KASR for Big Data Applications

Shakhy P S[1*] and Vidya K S[2]

[1*,2]Computer Science Department, Marian Engineering College, Kerala University, India

www.ijcseonline.org

Received: April /02/2015 Revised: April/11/2015 Accepted: April/23/2015 Published: April/30/ 2015

***Abstract*—** Service recommender systems are valuable tools for providing appropriate recommendations to users. In the last decade the rapid growth of the number of customers, services and other online information yields service recommender systems in Big Data environment, some critical challenges .Traditional service recommender systems often suffer from scalability and inefficiency problems when processing or analyzing such large scale data. Moreover, most of the existing service recommender systems present the same ratings and rankings of services to different users without considering diverse users' preferences, and therefore fails to meet users' personalized requirements . KASR(Keyword Aware Service Recommendation System) aims at calculating a personalized rating of each candidate service for a user by extracting keywords from user reviews, and then presenting a personalized service recommendation list and recommending the most appropriate services to users. Various limitations of the current recommendation methods can be reduced by possible extensions that can provide better recommendation capabilities. These extensions include incorporation of the contextual information into the recommendation process. Designing and implementing scalable recommender systems in Big Data environment solve the scalability problem.

Keywords— Keyword Aware Service Recommendation System , Collaborative Filtering, BigData

I. INTRODUCTION

Recommender systems are software applications that attempt to reduce information overload by recommending items of interest to end users based on their preferences .It can be defined as a system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful services in a large space of possible options. The first recommender system, Tapestry, was designed to recommend documents from newsgroups[11]. Nowadays, the trend "everything as a service" has been creating a Big Services era due to the foundational architecture of services computing and a way of offering social networking services, big data analytics, and Internet services. The big data comprises high volume, high velocity, and high variety information assets , which are difficult to gather, store, and process by using the available technologies.A keyword-aware service recommendation method, named KASR, uses a user-based Collaborative Filtering algorithm. In KASR, keywords extracted from reviews of previous users are used to indicate their preferences and to generate new recommendations.

II. RECOMMENDATION METHODS

Current recommendation methods can be usually classified into three main categories: content-based, collaborative, and hybrid recommendation approaches. . Content-based approaches recommend services similar to those the user preferred in the past. Collaborative filtering (CF) approaches recommend services to the user that users with similar tastes preferred in the past. Hybrid approaches combine content-based and CF methods in several different ways.

Content-based filtering

Content-based recommender systems work with profiles of users that are created at the beginning. A profile has information about a user and his taste. Taste is based on how the user rated items. Generally, when creating a profile, recommender systems make a survey, to get initial information about a user in order to avoid the new-user problem. The content-based approach to recommendation has its roots in information retrieval and information filtering . Because of the significant and early advancements made by the information retrieval and filtering communities and because of the importance of several text-based applications, many current content-based systems focus on recommending items or services containing textual information, such as documents, Web sites (URLs), and Usenet news messages. [2].

Collaborative filtering

Collaborative filtering became one of the most researched techniques of recommender The idea of collaborative filtering is finding users in a community that share appreciations . If two users have same or almost same rated items in common, then they have similar tastes. Such users build a group or a so called neighborhood. A user gets

recommendations to those items that he/she hasn't rated before, but that were already positively rated by users in his/her neighborhood [2].

In CF based systems, users receive recommendations based on people who have similar tastes and preferences, which can be further classified into item-based CF and user-based CF.In the user-based approach the items that were already rated by the user before, play an important role in searching a group that shares appreciations with him.In item-based systems, the predicted rating depends on the ratings of other similar items by the same user.

Hybrid recommendation approaches

For better results some recommender systems combine different techniques of collaborative approaches and content based approaches. Using hybrid approaches we can avoid some limitations and problems of pure recommender systems, like the cold-start problem.

.

III. MODERN RECOMMENDATION APPROACHES

Context-aware approaches

Context is the information about the environment of a user and the details of situation he/she is in. Such details may play much more significant role in recommendations than ratings of items, as the ratings alone don't have detailed information about under which circumstances they were given by users.Context-aware recommender systems became much attention, as they noticeably increased the quality of recommendations and the approaches became more specific to use in certain areas[7].

Semantic based approaches

Most of the descriptions of items, users in recommender systems and the rest of the web are presented in the web in a textual form. Using tags and keywords without any semantic meanings doesn't improve the accuracy of recommendations in all cases, as some keywords may be homonyms. Traditional text mining approaches that is based on lexical and syntactical analysis show descriptions that can be understood by a user but not a computer or a recommender system. That was a reason of creating new text mining techniques that were based on semantic analysis. Recommender systems with such techniques are called semantic based recommender systems[2].

Cross-domain based approaches

Finding similar users and building an accurate neighborhood is an important part of recommending process of collaborative recommender systems. Similarities of two users are discovered based on their appreciations of items.In cross-domain systems similarities of users computed domain-dependent. Recommender system determines the overall similarity, creates overall neighborhoods and makes predictions and recommendations[2].

Peer-to-Peer approaches

The recommender systems with P2P approaches are decentralized. Each peer can relate itself to a group of other peers with same interests and get recommendations from the users of that group. Recommendations can also be given based on the history of a peer. Decentralization of recommender system can solve the scalability problem.

Cross-lingual approaches

The recommender system based on cross-lingual approach lets the users receive recommendations to the items that have descriptions in languages they don't speak and understand .The main idea of cross-lingual based approach is to map both text and keywords in different languages into a single feature space, that is to say a probability distribution over latent topics. From the descriptions of items the system parses keywords than translates them in one defined language using dictionaries. After that, using collaborative or other filtering, the system gives recommendations to users .Cross-lingual recommender systems break the language barrier and gives opportunities to look for items, information, papers or books in other languages[2].

Keyword based approach

In this method, keywords are used to indicate both of users' preferences and the quality of candidate services. A user-based CF algorithm is adopted to generate appropriate recommendations. KASR(Keyword Aware Service Recommendation System) aims at calculating a personalized rating of each candidate service for a user, and then presenting a personalized service recommendation list and recommending the most appropriate services to him/her .Two data structures use din this method are , "keyword-candidate list" and "specialized domain thesaurus", introduced to help obtain users' preferences The keyword-candidate list is a set of keywords about users' preferences and multi-criteria of the candidate services, which can be denoted as $K=\{k1,k2,\ldots,kn\}$, n is the number of the key-words in the keyword-candidate list .The preferences of previous users will be extracted from their reviews for candidate services and formalized into a keyword set. A domain thesaurus is a reference work of the keyword-

candidate list that lists words grouped together according to the similarity of key-word meaning, including related and contrasting words and antonyms .FIG 1 shows a typical KASR system.

The main steps of KASR includes capturing preferences of an active user ,computing the similarity of active user preference to a previous user preference ,and recommending personalized services to the active user by keyword aware approach . An active user can give his/her preferences about candidate services by selecting keywords from a keyword-candidate list, which reflect the quality criteria of the services he/she is concerned about. The preference keyword set of the active user can be denoted as APK={ak1,ak2,..,ak*l*} , where *l* is the number of selected keywords . A review of the previous user will be formalized into the preference key-word set of him/her, which can be denoted as PPK={pk1,pk2,…,pk*h*},where *h* is the number of extracted keywords .Based on the similarity of the active user and previous users, further filtering will be conducted. 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(s) with the highest rating(s) will be recommended to him/her[1].

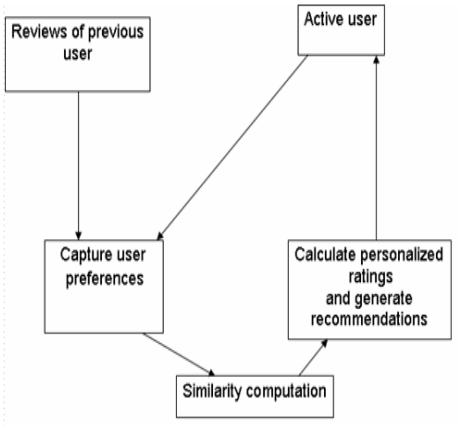

FIG 1: KASR SYSTEM

IV. SIMILARITY COMPUTATION

There are different algorithms of measuring similarities among items or services in data base and those in users profile .The Keyword based approach uses Approximate similarity computation and Exact similarity computation.

Approximate similarity computation

A frequently used method for comparing the similarity and diversity of sample sets, Jaccard coefficient, is applied in the approximate similarity computation. Jaccard coefficient is measurement of asymmetric information on binary (and non-binary) variables, and it is useful when negative values give no information. The similarity between the preferences of the active user and a previous user based on Jaccard coefficient is described as follows

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

where *APK* is the preference keyword set of the active user, *PPK* is the preference keyword set of a previous user.

Exact similarity computation

A cosine-based approach is applied in the exact similarity computation, which is similar to the Vector Space Model (VSM) in information retrieval. The preference keyword sets of the active user and previous users will be transformed into *n*-dimensional weight vectors respectively .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[1].

Rating products and services

Popularity of an item indicates how frequently users rated the item. Popular items may be good at connecting people with each other as co-raters, since many people are likely to rate popular items. A weighted average approach can be used to calculate the personalized rating of a service for the active user. Personalized ratings of all candidate services for the active user can be calculated in the same manner. Then rank the services by the personalized ratings and present a personalized service recommendation list to the active user .

V. TOOLS AND TECHNIQUES

Hadoop

In order to manage and structure data, Hadoop is emerging as a core platform for big data applications. Hadoop implements a computational paradigm named MapReduce, where data and application are divided into small fragments and distributed among different nodes.

Mahout

Apache Mahout machine learning library is written in Java that is designed to be scalable, i.e. run over very large data sets. It achieves this by ensuring that most of its algorithms are parallelizable (map-reduce paradigm on Hadoop. The Mahout project was started by people

involved in the Apache Lucene project with an active interest in machine learning and a desire for robust, well-documented, scalable implementations of common machine-learning algorithms for clustering and categorization.

MapReduce

Similar to most big data applications, the big data tendency also poses heavy impacts on service recommender systems. To improve its scalability and efficiency in big data environment, KASR is implemented on Hadoop, a widely-adopted distributed computing platform using the MapReduce parallel processing paradigm. Hadoop is the most popular open source cloud computing platform inspired by MapReduce and Google File System ,which supports MapReduce programming framework and mass data storage with good fault tolerance. MapReduce is a popular distributed implementation model proposed by Google. MapReduce is a programming model and an associated implementation for processing and generating large data sets.

VI. DIFFICULTIES WITH RECOMMENDATION SYSTEMS

Cold-start

Its difficult to give recommendations to new users as his profile is almost empty and he hasn't rated any items or services yet so his taste is unknown to the system. This is called the cold start problem. In some recommender systems this problem is solved with a survey when creating a profile. Both of these problems can be also solved with hybrid approaches.

Scalability

With the growth of numbers of users and items/services, the system needs more resources for processing information and forming recommendations. To address the scalability issues, a few proposals applied model based CF. The model-based approaches apply data mining and machine learning algorithms to find patterns based on the training data to reduce the size of the user-item rating matrix .

Sparsity

In online shops that have a huge amount of users and items there are almost always users that have rated just a few items/services Sparsity is the problem of lack of information. In any recommender system, the number of ratings already obtained is usually very small compared to the number of ratings that need to be predicted.

One way to overcome the problem of rating sparsity is to use user profile information when calculating user similarity. Limited content analysis ,privacy, and overspecialization are the other problems. Table 1 shows the comparison of existing recommendation approaches highlighting the disadvantages.

Recommendation Approach	Technique Used	Disadvantages
Content based	Information Retrieval & filtering (TF-IDF)	Limited content analysis, new user problem, Overspecialization
Collaborative	Nearest neighbor	New user & new item problem ,Sparsity
Hybrid	Combine content based & collaborative	Sometimes quality & accuracy gets affected
Context Aware	Includes information about user environment	Changes dynamically ,periodical refreshing is needed
Keyword Aware	Extracts keywords from user reviews	Intrusive in nature

TABLE1: Comparison of existing recommendation approaches

VII. CONCLUSION

Nowadays, recommendation systems are increasingly gaining notoriety due to their height number of applications. The users cannot manage all the information available on Internet so it is necessary to provide recommendations.. Recommender systems made significant progress over the last decade when numerous content-based, collaborative, and hybrid methods were proposed and several "industrial- strength" systems have been developed. However, despite all of these advances, the current generation of recommender systems surveyed still requires further improvements to make recommendation methods more effective in a broader range of applications. Various limitations of the current recommendation can be reduced by possible extensions that can provide better recommendation capabilities .

FUTURE WORK

For improving the efficiency of KASR in Big Data environment and to decrease the intrusiveness some contextual information and time constraints can also be considered along with the keywords. Incorporation of time further reduce the amount of data being processed as fresh reviews are considered .Designing and implementing scalable recommender systems in "Big Data" environment solve the scalability problem by dividing the datasets This significantly improves the accuracy and scalability of service recommender systems over existing approaches and performs better with larger datasets.

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] Gediminas Adomavicius, and Alexander Tuzhilin "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions" IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, June 2005.

[3] 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.

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

[5] J. Dean, and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Communications of the ACM, Vol. 51, No.1, pp. 107-113, 2005.

[6] Daniar Asanov "Algorithms and Methods in Recommender Systems" Berlin Institute of Technology Berlin, Germany.

[7]Osman Khalid Muhammed Usman Shahid Khan, Samee U Khan , Albert Y Zomaya "Omnisuggest: A Ubiquitous Cloud Based Context Aware Recommendation System for Mobile Social Networks " IEEE Transactions on Services Computing ,2013.

[8] Sang Hyun Choi, Young-Seon Jeong, and Myong K. Jeong "A Hybrid Recommendation Method with Reduced Data for Large-Scale Application" IEEE Transactions on systems, man, and cybernetics—Part C: Applications and Reviews, Vol. 40, No. 5, September 2010

[9] Building Recommendation Platforms with Hadoop, Jayant Shekhar O'Reilly Strata Conference Making Data Work.

[10] "Recommender Systems Handbook" Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol Springer Science+Business Media, LLC 2011,pp 39-67.

[11] Recommender Systems , Daniel Rodriguez, University of Alcala ,Windsor ,August 5,2013