International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 09 Issue 11 ǁ 2021 ǁ PP. 29-35

Hadoop Map Reduce Framework for Keyword-Aware Recommendation System with Sentiment Analysis

Shireen Khan
M. Tech. Scholar, All Saints' College of Technology, Bhopal, India

Sarwesh Site
Professor, Dept. of CSE, All Saints' College of Technology, Bhopal, India

ABSTRACT*: Recommender systems apply knowledge discovery techniques to the problem of makingpersonalized product recommendations using customers usage pattern. Service recommendersystems provide an important tool for information gathering, information filtering, informationprocessing, and recommending services to the users with appropriate results. The regularly andrapidly growing number of users, services, and data are presenting new challenges. Newrecommender systems technologies are needed to quickly pro-duce high qualityrecommendations, even for very large-scale problems. There exist a lot of recommendation*
methods currently but these existing recommender systems needs to be improved, efficient, andtrustworthy to meet the present requirements. To address these challenges, we propose aKeyword-Aware Service Recommendation method with Sentiment Analysis to recommendtrustworthy services to the user. This research is based on the work done in a article named Keyword-Aware ServiceRecommendation method (KASR) on MapReduce for Big Data Applications. KASR is basedon Preference-Aware Service Recommendation method (PASR) and it uses a user-basedcollaborative filtering algorithm that searches for specific keywords. KASR is implemented onHadoop environment to handle the big data generated by recommendation systems to improvethe scalability and efficiency. KASR considers user's preferences but lacks the sentiment
analysis that may result in an effective number of wrong recommendations. We proposed thekeyword aware recommendation system with sentiment analysis algorithm. We addressperformance issues by implementing the algorithm using Hadoop Map-Reduce frameworkcombined with similarity based collaborative filtering.
KEYWORDS*: Sentiment Analysis,Hadoop,KASR,PASR,Machine Learning.*

Date of Submission: 12-11-2021 Date of acceptance: 28-11-2021

I. INTRODUCTION

Recommendation systems have turned out to be greatly useful for number of practical applications in recent years, and these applications have wide domain of variety. Films, music, news, research articles, e-marketing, social networking, and consumer based online productionsare very common applications where a recommendation system is a need. Generally, recommendation systems utilizes the response of previous users who purchased or used andreviewed products to provide suggestions for current or future users. These systems take thereviews of past customers similar to current customer and usually current preferences of thecustomer as keywords, then aggregates them on the basis of certain rules to provide recommendations as output.

Due to heavily increasing amount of data, these algorithms needto process intensive large information sets. Now days applications like social networking websites and search engines generate several gigabytes to several terabytes and even up toseveral petabytes of data [1][2]. Several years ago Google published a paper presenting theMapReduce framework to deal with its ever growing enormous amount of data in distributedenvironment. MapReduce framework simplifies the complexity of processing large amount ofinformation stored on distributed node clusters in parallel manner, because it facilitatesprogrammers so that they can map a instruction among different nodes in the cluster to executeon different set of data. MapReduce deals with the distributed environment and parallelism omits own, reducing the complexity for programmers, and results in minimalistic required data[3].

Self data replication in MapReduce model offers great fault tolerance also[4]. MapReduce dealswith data distribution, parallel processing, and big data in quite efficient way. Many majorcompanies have implemented their own MapReduce models with their own set of requirementsInformation filtering algorithms are well known for prediction of preferences or ratings thatuser would give to a product. Recommender algorithms are subset of information filteringalgorithms. Most of these filtering algorithms are static in nature providing a generic experienceto every user, meaning users just search and buy products. But recommendation systemsprovide greater user interaction to seller, and richer experience to users. To provide effectiverecommendations, recommender systems take the advantage of old data such as past records ofpurchases, searches, reviews, and users' behavior. Correct implementation of recommendersystems can be extremely effective at increasing user engagement and purchasing. Today,many of the worlds' most heavily trafficked websites, such as Google, Facebook, LinkedIn,Amazon, and Twitter employ recommendation systems to engage their users with appropriateand personalized content[5][6]. Figure 1 show the Recommendation System Process

Figure1.Framework for Keyword-Aware Recommendation System

To solve the recommendation problem the concept of knowledge data discovery is used bytaking in users previous responses in the form of purchased items, user's behavior on websites,user feedbacks, and ratings given by users. Recommendation systems enhance the userexperience by providing suggestions to users such as what products to buy, which article to read, and which movie to watch. In research field many different approaches have been madeto address the problem of making efficient and accurate personalized recommendations. Someof these approaches are Data mining, machine learning,Collaborative Filtering, User based,Item based[7][8].

In most traditional recommender systems, users are either presented with a preferredrecommendations or a user can give some textual or selection inputs (user's preferences)[9] tosort the list out. In some modern recommender systems, users are recommended using previoususer's selections and feedback along with the user's preferences [10]. So in the first case theyare not considering the user's preferences. Users are not same, so a good system should considerthe users requirement. While in the second case they are not considering the sentiments of thefeedback provided by the previous users. They are only taking it as positive feedback and notconsidering negative feedback, which may provide wrong recommendations as in their methodthey are filtering previous users' feedbacks with the keyword entered by the current user.

Thismay not work in case of sentences with negative sense means negative feedback. In this research, keyword aware service recommendation systems are researched broadly and a newapproach for keyword-based recommendation system is proposed to provide betterrecommendations.

Motivated by these observations, in this paper, we address these problems by introducing someimprovement in one pre-existing recommendation approach, named as "Keyword-AwareService Recommendation Method on MapReduce for Big Data Applications".

The Thesis report is organized into 5 sections. Section 2 gives the literature review byexplaining the major research and papers about keyword-aware recommendation systems. In Section 3, we present the design and implementation of the proposed algorithm.In Section 4, we provide the experiments along with the results of this algorithmic approach.We discuss the future work that can be done in this research and concludes our work in Section 5.

II. RELATED WORK

The First recommender system was developed by Goldberg, Nichols, Oki Terry[11] in 1992.Tapestry was an electronic messaging system that allowed users to either rate messages ("good"or "bad"). Recommender system as defined by M. Deshpande and G. Karypis: A personalizedinformation filtering technology used to either predict whether a particular user will like aparticular item (prediction problem) or to identify a set of N items that will be of interest to acertain User. Recommender systems form or work from a specific type of information filteringsystem technique that attempts to recommend information items (movies, TVprogram/show/episode, video on demand, music, books, news, images, web pages, scientificliterature etc.) or social elements (e.g. people, events or groups) that are likely to be of interestto the user. Typically, a recommendersystem compares a user profile to some reference

characteristics, and seeks to predict the 'rating' or 'preference' that a user would give to an itemthey had not yet considered.

Shunmei Meng in 2014 proposed a KASR[12] method for personalized recommendation. Inthis user based collaborative filtering is used. For more efficiency the method is implementedon Hadoop. For evaluation Jaccard coefficient and Cosine similarity measure is used. User'spositive and negative reviews are not considered separately. Sentiments in the reviews are notconsidered. To make the method more efficient and scalable Hadoop MapReduce model is used.

Xiwang yang in 2013 proposed a Bayesian inference based recommendation in online socialnetworks[13]. In this users share their content ratings with friends. Rating similarity is measured using conditional probability. Based on similarity score ranking and recommendation is done.

There is a Cold start and rating sparseness problem.In [14], the authors propose a Bayesian inference based recommendation system for online social networks. They show that the proposed Bayesian inference based recommendation isbetter than the existing trust based recommendations and is comparable to CollaborativeFiltering recommendation.

In [15], Adomavicius and Tuzhilin give an overview of the field of recommender systems anddescribe the current generation of recommendation methods. They also describe variouslimitations of current service recommendation methods, and discuss possible extensions thatcan improve recommendation capabilities and make recommender systems applicable to aneven broader range of applications. Most existing service recommender systems are only basedon a single numerical rating to represent a service's utility as a whole. In fact, evaluating aservice through multiple criteria and taking into account of user feedback can help to makemore effective recommendations for the users.

Guoshengkang[16] proposed an active web service recommendation. Web usagehistory and QoS are the main criteria for recommendation. Using this recommendation top kservices are generated for users. Passive users reviews are not considered. Using this approachtop k services are generated for users. Passive users reviews about the website is not considered..

Yan Ying Chen[17] proposed a probabilistic personalized travel recommendation model. People attributes and photos are used which are effective for mining demographics fortravel landmarks and paths, and thus greatly benefiting personalized travel recommendation.

FaustunoSanchez[18] proposed a recommender system for sport videos, transmitted over the Internet and/or broadcast, in the context of large-scale events, which has been testedfor the Olympic Games.

ZibinZheng[19] proposed quality of service ranking prediction for cloud services. Thispaper investigate the combination of rating based approaches and ranking based approaches, sothat the users can obtain QoS ranking prediction as well as detailed QoS value prediction. Howto detect and exclude malicious QoS values provided by users is not proposed here.With the development of cloud computing software tools such as Apache Hadoop, MapReduce,and Mahout, it becomes possible to design and implement scalable recommender systems in

“Big Data” environment.[20] The authors of implement a CF algorithm on Hadoop. They solve

the scalability problem by dividing dataset. But their method doesn't have favorable scalabilityand efficiency if the amount of data grows.

Jin et al. [21] propose a large-scalevideo recommendation system based on an item-based CF algorithm. They implement their proposedapproach in Quist, which is a .Net MapReduce framework, thus their system can work for largescale video sites.

III.PROPOSED METHDOLOGY

This chapter explains the proposed design used in the implementing keyword aware servicerecommendation system with sentiment analysis on MapReduce.Figure 2 shows the basic framework for key-aware recommendation system with sentiment analysis on MapReduce.

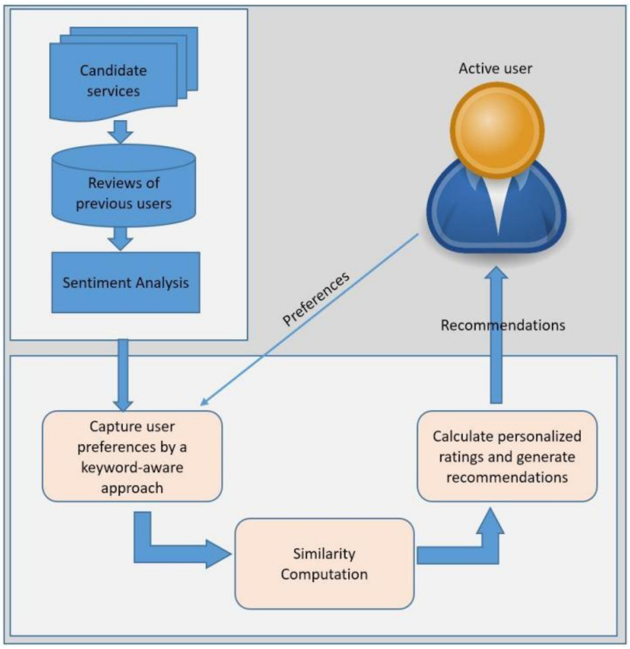

Figure2:**Framework of Keyword-Aware Recommendation System with Sentiment Analysis**

Capturing User's Preferences

There are two types of users: Active user and Previous users. Active user is the user which iscurrently using the services by providing set of keywords according to his preferences from theKeyword List. Previous users are the users which has already used the services and reviewedfor the services. Based on the reviews from the previous user, the keywords are extracted fromthe reviews according to Keyword List and Domain Thesaurus.

Keyword Extraction

In this phase, each review will be transformed into a corresponding keyword set according tothe keyword-candidate list and domain thesaurus. If the review contains a word in the domainthesaurus, then the corresponding keyword should be extracted into the preference keyword setof the user. For example, if a review of a previous user for a hotel has the word "spa", which iscorresponding to the keyword "Fitness" in the domain thesaurus, then the keyword "Fitness"should be contained in the preference keyword set of the previous user. If a keyword appearsmore than once in a review, the times of repetitions will be recorded. In this paper, it is regardedthat keywords appearing multiple times are more important. The times of repetitions will beused to calculate the weight of the keyword in preference keyword set in the next step.

Similarity Computation

The second step is to identify the reviews of previous users who have similar tastes to an activeuser by finding neighborhoods of the active user based on the similarity of their preferences.Before similarity computation, the reviews unrelated to the active user's preferences will befiltered out by the intersection concept in set theory. If the intersection of the preferencekeyword sets of the active user and a previous user is an empty set, then the preference keywordset of the previous user will be filtered out.

Approximate Similarity Computation

A frequently used method for comparing the similarity and diversity of sample sets, Jaccardcoefficient, is applied in the approximate similarity computation. Jaccard coefficient is measurement of asymmetric information on binary (and non-binary) variables, and it is usefulwhen 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: The formula usedby Jaccard Coefficient is

$$simASC(APK, PPKj) = \frac{|APK \cap PPKj|}{|APK \cup PPKj|}$$

Where APK is the preference keyword set of the active user, PPK is the preference keyword setof a previous user. And the weight of the keywords is not considered in this approach.

Exact Similarity Computation

In Exact Similarity Computation, the similarity between preference of active user and previoususers is calculated based on Cosine-Based approach .The formula used by Cosine-Basedapproach is

$$simESC(APK, PPK) = cos(WAP, WPP) = \frac{\text{WAP.WPP}}{||WAP||2 \times ||WPP||2}$$

Preference weight vector: In this cosine based approach, the preference keyword sets of activeuser and previous users is transformed into n-dimensional weight vectors respectively, namelypreference weight vector. The preference weight vectors of the active user and a previous userare noted as WAP and WPP, respectively. In our research, we use the Analytic Hierarchy Process (AHP) model to decide the weight ofthe keywords in the preference keyword set of the active user. AHP method is provided bySaaty in 1970s to choose the best satisfied business role for its hierarchy nature. The weightcomputing based on the AHP model is decided as follows:

Firstly, we construct the pair-wise comparison matrix in terms of the relative importancebetween each two keywords. The pair-wise comparison matrix Am = (aij)m must satisfy thefollowing properties, aij represents the relative importance of two keywords:

aij = 1, i = j = 1, 2, 3, …,m.

aij = 1/aji ,i, j = 1, 2, 3, ….,m and i ≠ j.

aij = aik / ajk ,i, j, k = 1, 2, 3, …, m and i ≠j.35

After checking the consistence of the matrix, then we calculate the weight by the following function:

$$wi = 1 / m\sum_{j=1}^{m} \frac{aij}{\sum_{k=1}^{m} akj}$$

Here, aij is the relative importance between two keywords; m is the number of keywords in thepreference keyword set of active users.

IV .RESULT ANALYSIS

The algorithm is designed to derive recommendations to the users based on their preferencekeywords and reviews of previous users. Experiments are done on both single node and multinode clusters with varying data sizes and performance metrics are recorded. All the map-reducejobs are run in sequence as the output of the previous jobs are given as input for the next job

Provided experiment results are based on the comparison between execution times taken byKASR_ESC algorithm and proposed approach KASRwithSentiAnalysis.

Figure 3 shows the comparative analysis of both the algorithms on 250MB data set in singlenode cluster, 2-node cluster, 4-node cluster, and 6-node cluster. From the graph it is clear thatKASRwithSentiAnalysis is performing better in single node cluster and 2-node cluster. But in4-node cluster and 6-node cluster very slight improvement in proposed algorithm.

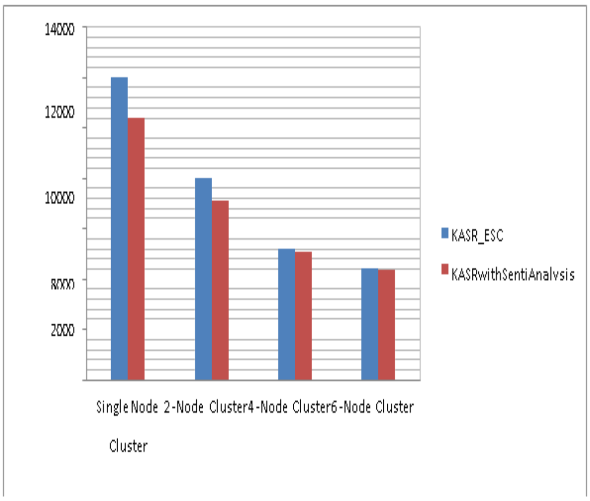

Figure 3:- Comparative Analysis of KASR_ESC and KASRwithSentiAnalysis on 250MB data set

Figure 4 shows the comparative analysis on 1GB data set. We can see that as the data size increases the performance of KASRwithSentiAnalysis increases or atleast same as of KASR in single node and 2-node cluster. is performing better in single node cluster and 2-node cluster

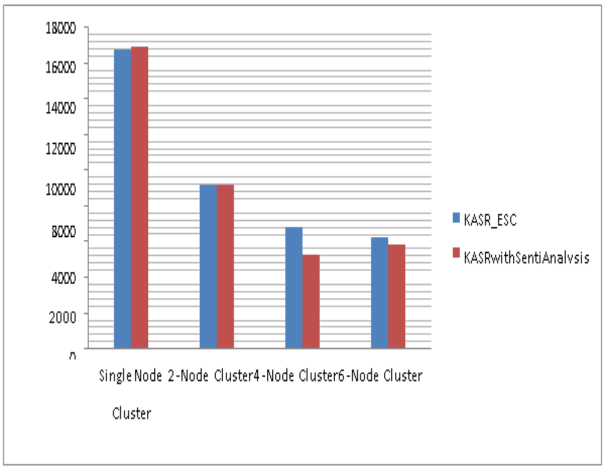

Figure 4 Comparative Analysis of KASR_ESC and KASRwithSentiAnalysis on 1GB data set

Next we experimented both the algorithms on 10GB data set. As the data size increases more, the performance of proposed algorithm degrades. This is due to the overhead of the computation for sentiment analysis, but also gives better result than KASR.

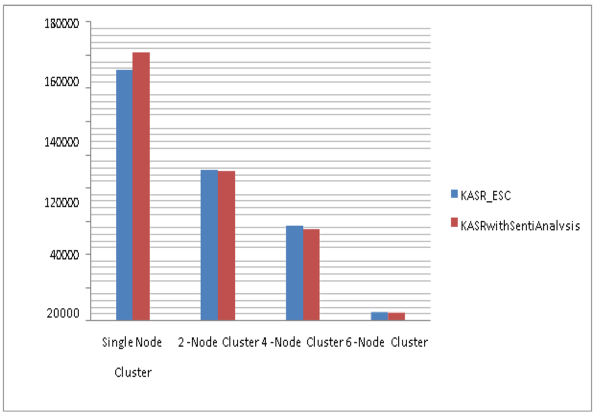

Figure 5. Comparative Analysis of KASR_ESC and KASRwithSentiAnalysis on 10GB data set

V. CONCLUSION

In this paper we discussed about recommendation system algorithms and MapReduceprogramming model by Hadoop. A sequential algorithm to compute similarity is explained. Then we proposed keyword-aware recommendation system with sentiment analysis to providebetter recommendations to users. Our approach provide recommendations by taking the userpreferences and utilizing the reviews given by previous users. The major improvement was toanalyze the sentiments of users by processing their reviews. Quality wise we achieved betterresults than the original algorithm.

Different experiments are conducted to show the performance improvement of using a multinode cluster. Test data for experiments is taken from data bases of a hotel review system.

The Hadoop map-reduce approach saves a lot of resources incomputing similarities and generates recommendations in short time.

For future work, lot of improvements can still be done. Sentiment analysis algorithm can bemore accurate and optimized. Another better approaches can be applied to identify thesentimental behavior more accurately

REFERENCES

[1]. Cheikh Kacfah Emani, Nadine Cullot and Christophe Nicolle "Understandable Big Data: A survey" in Computer Science Review Volume 17, August 2015, Pages 70–81.

[2]. H. Hu, Y. Wen, T.-S. Chua, and X. Li, ''Towards scalable systems for big data analytics: A technology tutorial,'' IEEE Access, vol. 2, pp. 652–687, 2014.

[3]. Rodrigo Agerri,XabierArtola, ZuhaitzBeloki, German Rigau, AitorSoroa "Big data for Natural Language Processing: A streaming approach" in Knowledge-Based SystemsVolume 79, May 2015, Pages 36–42.

[4]. A. Rabkin and R. H. Katz, "How Hadoop Clusters Break," IEEE Software, vol. 30, pp. 88- 94, 2013.

[5]. T. Jiang, Q. Zhang, R. Hou, L. Chai, S. A. McKee, Z. Jia, and N. Sun, "Understanding the behavior of in-memory computing workloads," in Workload Characterization (IISWC), IEEE International Symposium on, 2014, pp. 22–30.

[6]. Martin Zwilling. What Can Big Data Ever Tell Us About Human Behavior [online]. Available: http://www.forbes.com/sites/martinzwilling/2015/03/24/what-can-big-data- ever-tell-us-about-human-behavior/#1580346f1bed. Date accessed: (March 24, 2015).
[7]. Saeed Shahrivari, "Beyond Batch Processing: Towards Real-Time and Streaming Big Data", Computers, Vol. 3, pp. 117.129, 2014.
[8]. Sachin Agarwal. Monitoring and Troubleshooting Apache Storm [online]. Available: https://dzone.com/articles/monitoring-and-troubleshooting-apache-storm-with-o. Date accessed: (April 7, 2016).
[9]. M. R. Evans, D. Oliver, K. Yang, X. Zhou, S. Shekhar, "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities," Ed. S. Wang, M. F. Goodchild, CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery. Springer, 2014.
[10]. Wikipedia. "Spatialdatabase"[online]. Availablehttps://en.wikipedia.org/wiki/Spatial_database.
[11]. Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen," KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications" IEEETransactions On Parallel And Distributed Systems, TPDS-2013-12-11.
[12]. 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
[13]. A. Tuzhilin and G. Adomavicius, "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, pp. 734-749, 2005.
[14]. 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.
[15]. 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.
[16]. 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.
[17]. Zibin Zheng, Xinmiao Wu, YileiZhang,Michael R. Lyu, Fellow,andJianmin Wang," QoS Ranking Prediction for Cloud Services" IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 6, June 2013.
[18]. 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.
[19]. Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (2012). GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 40(3), pp. 77-87.. [Online; accessed June 2014].
[20]. Adomavicius G., Tuzhilin A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.. [Online; accessed De- ecember 2014].
[21]. A. Rabkin and R. H. Katz, "How Hadoop Clusters Break," IEEE Software, vol. 30, pp. 88- 94, 2013.