Matrix Academic International Online Journal Of Engineering And Technology
ISSN: 2348-3326, SJIF 2013 3.9
www.maioj.org © MAIOJET
Volume 4, Issue 1

Toward personalize user-based recommendation system for big data application.

Nasim Kothiwale[#1], Prof. Shrihari Khtawkar [*2]

[#]*Computer Engineering, Shivaji University*
Kolhapur, Maharahstra,India.

[1]nasim22nov@gmail.com

[2]sdk_cse@adcet.in

***Abstract*—The use of web are increases the option of user choices also increases. Recommendation system is guideline for users in personalized way to choose/choice their option from the large space datasets. Large volume dataset is refers to "Big Data", Big data is nothing but the data which is beyond the capacity of storing, managing and processing within a short time period. In this paper purposes the personalized user-based recommendation system in which the movies recommendation list is generated as per user interest. In previous service recommendation system collaborative filtering algorithm is adopted but they are faces problem with scalability and inefficiency at the time of data retrieval. The existing service recommendation systems are fails to meet user requirements because without considering users preferences/interest's it display same ratings and rankings to different users. Also in traditional recommendation system yielding the big data discovery and analysis problem. In purposed system ratings or features are used to filtering the information by applying multi criteria selection policy. Basically to manage and solve scalability and efficiency problem Hadoop is broadly adopted distributed computing platform with MapReduce parallel processing environment. Finally, Experiment is conducted on real-world data set and results demonstrate the accuracy, efficiency and scalability to improve recommendations.**

***Keywords*—Recommendation system (RS), Collaborative filtering (CF), Hadoop, MapReduce, Big Data, Ratings.**

I. INTRODUCTION

The web is leaving the era of search and entering one of discovery. Search is what you do when you're looking for something where discovery is when something wonderful that you want. With the rapid development of latest technologies simultaneously web 2.0 eras is started. Beginning of this era with different opportunities such as speed in processing, sharing information, opining of other users etc. comes out in consideration.

As the use of internet increases people may come with multiple choices where to travel? Which book to buy? Which movie to watch? And so on. Therefor researches need a powerful technique to extract knowledgeable and useful data from the large volume of data. Data comes from the several data sources such as social networking sites, online transaction, scientific research areas, sensor technology and network, data sharing and so on. To handle and manage this huge amount of data some challenges are found those are discovery of data, scaling problem, mining complex data, analysis of useful data, and extraction of knowledgeable data etc. at that time Bid Data comes under consideration.

Big data is the term for collection of data sets so it is large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing application.

Recommendation system play an important role to addressed all above problems. It is solution against Big Data handling and management. Recommendation system (RS) is a technique for filtering and sorting services as per their features and ratings to generate the more accurate recommendation list. RS is live interaction and producing high quality recommendation.

A RS provide useful and effective suggestions for specific types of item it's normally focuses on a design, and core recommendation techniques. There are several approaches are introduced but the personalized user-based is most effective one. It recommends items as per user interest.

Recommendation system works on the collection and processing large-space data "Big Data". For processing purpose MapReducer are used which is parallel processing module of hadoop. Hadoop is platform for distributed computing.

II. MOTIVATION

In the past few years the amount online data speedily goes increases in the size. This yields success of web 2.0. Several organization works on large scale information about their employee, customer, services, provider and operations.

The existing service recommendation system posture critical challenges when data frequently goes increases. They are presented some rankings and services recommendation list. Without consider user's personal requirement or interest.

The most recommendation system uses collaborative filtering algorithm. The problem with CF is Data Sparsity, Extendibility, and Similarity.

As the success of web 2.0 the data extremely increases which exuded the capacity of managing, handling and processing within short period of time.

The traditional recommendation system are poses the big data analysis and discovery problem. Also it is often suffers

from accuracy, speedup (low processing speed), and efficiency problem when working on bulky data in parallel and distributed system environments.

Motivated by these observation in purposed system presents personalize user-based recommendation system list for movie which will produce accurate recommendation list as per user interest.

III. RELATED WORK

The authors Shunmei Meng, Wanchun[1] presents personalize recommendation list and as per user interest recommend the most appropriate items to the users. In keyword based recommendation system collaborative filtering algorithm is implemented on hadoop to raise appropriate recommendation. Hadoop is used for to improve scalability and efficiency.

The authors, X. Yang, Y. Guo, and Y. Liu [2], implemented Bayesian-inference-based recommendation system which is used for social networking sites. This is excellent than existing trust-based recommendation system. They show that active user gives reviews to every item and these reviews can see all users which are connected to active user.

In [3], Adomavicius and Tuzhilin give an overall structure of recommender systems. It describes the current generation recommendation approaches. In next generation recommendation system gives the solution on limitation of current generation recommendation system. This recommendation system gives idea about how to improve and how to make robust system. It is flexible in even broader range of application.

The [4] existing recommendation system only single rating criterion is applicable to generate recommendation list. But if without considering single one criteria, considering multi-criteria is most effective than other. Here rating, features and attribute of items are used to produce most appropriate recommendation. The feedbacks of users are also tack into account at the time of recommendation.

The authors Z D Zhao and M. S. Shang of [5] developed CF algorithm on Hadoop parallel processing paradigm. This is favourable to scale the large application by dividing the datasets to solve the scalability problem. The MapReduce and cascading technologies are used to implement scalable recommendation also in falksonomy information if presents a parallel user profiling approaches.

M. Hu, H. Singh, D. Rule, M. Berlyant, and Z. Xie Y. Jin [6] presenting a large scale video recommendation system implemented by using item-based CF algorithm. They implement their proposed approach in Qizmt, which is a .Net MapReduce framework, thus their system can work for large scale video sites.

The authors [7] proposed a trust-aware system for generating personalized user recommendations in social networks. Its foundations lie on a reputation mechanism that is mathematically formulated, comprising both local and collaborative rating formation. The proposed system provides users with personalized positive and/or negative recommendations that can be used to establish new trust/distrust connections in the social network.

The author [8] proposed location-aware recommender system they introduces some special types of ratings those are spatial ratings and non- spatial ratings. Those techniques are efficient, scalable and accurate one for recommendation system as compared with traditional recommendation system. They deals with scalability problem solve by applying those technology on Hadoop.

In this paper author [16] presents personalized recommendations are used to support the activities of learners in personal learning environments and this technology can deliver suitable learning resources to learners. This paper models the dynamic multi-preferences of learners using the multidimensional attributes of resource and learner ratings by using data mining technology to alleviate sparsity and cold-start problems and increase the diversity of the recommendation list. The proposed method outperforms current algorithms on accuracy measures and can alleviate cold-start and sparsity problems and also generate a more diverse recommendation list.

IV. METHODOLOGY

In proposed novel method of personalize user-based recommendation system.

The steps in purposed system are:

- Pre-processing on Reviews.
- Prepare keyword preference list.
- Keyword Extraction.
- Calculate user preference/choices.
- Semantic analysis.
- Neighbourhood similarity calculation.
- Generate personalize Top-K movies recommendation list.

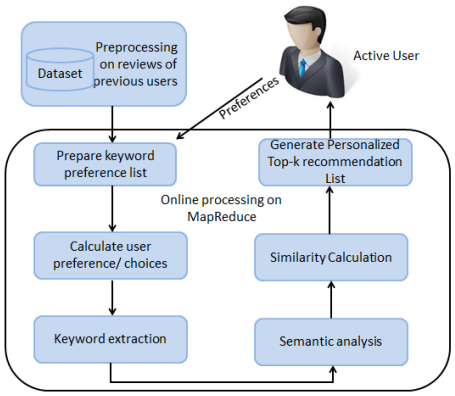

Fig. 1 System Architecture.

A. Pre-processing on Reviews

The pre-processing is done on the reviews of previous users. To filter the Dataset which contain the noise data? In pre-processing noise data is removed from the dataset. The noise data contain the same rating are for the multiple items or services. Also it contains unrated items and services. To remove the same rated and unrated data from the dataset is mandatory.

To done this we are using the Dynamic Noise Level (DNL) Algorithm.

B. Prepare keyword preference list

The preference list is a set of words about user's preference and multi-criteria of the services which can be denoted as k = {k1, k2,…,kn}, n is the number of the words in the preference list.

The user-based collaborative filtering algorithm is used for implement these one.

The algorithm summarizes the following steps:

1. All active users are weighted with respect to similarity. Similarity between active users is measured as the Pearson correlation between their ratings vector.
2. Select n active users that have the highest similarity.
3. Compute a prediction Pa,u from a weighted combination. Similarity between two users is computed using the person correlation coefficient.

$$Pa,u = \frac{\sum_{i=1}^{rn}(ra,j - ra)x(ru,j - ru)}{\sqrt{\sum_{i=1}^{rn}(ra,j - ra)^3}\ x\ \sqrt{\sum_{i=1}^{rn}(ru,j - ru)^3}}$$

Where r, a.i is the rating given to item I by user a, and ra is the mean rating given by user a.

TABLE 1
SAMPLE KEYWORD-PREFERENCES LIST OF MOVIE RECOMMENDATION

No	keyword	No	keyword
1	Action	6	Crime
2	Adventure	7	Drama
3	Animation	8	Horror
4	Children's	9	Romance
5	Comedy	10	Mystery

C. Calculate user preferences/choices

To the formalized into set of services the preference set of services. The preferences are extracted from the preference of previous users it compare with the preference of active users.

The peter stemmer algorithm is used for this method. A stemmer should conflict together all and only those pair of words which are semantically equivalent and share the same stem. Stemming is used to reduce the overhead of indexing and improve the performance of an IR system.

After extraction of combined preferences services which presenting in both, active user as well as pervious users services are displayed on users screen.

The algorithm summarizes the following steps:

1. Extract words from the keyword set.
2. Stem the words.
3. Calculate root word.
4. Display the preference list according to the root word.

$$sim(APL, PPL) = stemme(APL, PPL) = \frac{stm(APL)}{stm(PPL)}$$

D. Keyword extraction

In this phase, each review is transformed into corresponding keyword set according to preference list. If the reviews contain a keyword which is referencing to another keyword should be extracted from the set of keywords and generate the accurate preference list. Set of preference keyword are denoted by PKi = {pk1, pk2… pk n} where PKi ($1 \leq i \leq l$) is the keyword selected from active user, l is the number of selected words.

Basically here the degree of keyword also important.

TABLE II
DEGREE OF KEYWORDS

Degree of keyword				
General		Important	Very important	
1	2	3	4	5

The times of repetition is recorded which used to calculate the weight of words in preference list set for next step.

E. Semantic analysis

Semantic analysis is study of meaning of the words. The positive and negative meanings of words are considered. Here A={a1,a2,…an} are set of active users. N is no. of users, ai Є A. B={b1,b2,…,bn} are set of items and T= {t1,t2,…tn} are set of partial trust functions.

$$ti(aj) = \begin{cases} p, & if\ trust\ (ai, aj) = p \\ 1, & if\ no\ trust\ for\ aj\ from\ ai \end{cases}$$

R={r1,r2,…rn}are set of partial rating functions. t1Є T, foe every ai Є A.

$$ri(bj) = \begin{cases} p, & if\ rates\ (ai, bj) = p \\ 1, & if\ no\ ratings\ for\ bj\ from\ ai \end{cases}$$

F. Neighbourhood similarity calculation

Two neighbourhood similarity calculations are introduced in purposed recommendation system.

1. Approximate similarity calculation method.
2. Exact similarity computation method.

1. Approximate similarity calculation method.

The approximate calculation method is for the case the weights of the keywords in the preference set are unavailable. jaccard coefficient is measurement of asymmetric information on binary(and non-binary) variables, and is useful when negative values given no information. The similarity between the preference of the active user and previous user based on Jaccard coefficient is described as follows:

$$sim(APL, PPL) = jaccard\,(APL, PPL) = \frac{|APL \cap PPL|}{|APL \cup PPL|}$$

Where APL is a set of the preference list of active user, PPL is set of preference list of previous user.

2. Exact similarity calculation method.

The exact similarity calculation method is for the case that the weights of the keywords are available. a cosine-based approach is applied in the exact similarity computation, which is similar to the vector space model (VSM) in information retrieval.

Calculate the weight by the following function.

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

G. Generate personalized Top-k movies recommendation list.

Based on the similarity of the active user and previous users, further filtering will be conducted. Given a threshold δ, of sim(APK, PPKj) < δ, , the preference keyword set of a previous user PPKj will be filtered out, otherwise PPKj will be retained. 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(s) with the highest rating(s) will be recommended to him/her.

V. Result Analysis

Experiments are conducted to evaluate the accuracy and scalability of our recommendation list.

- *Movie Dataset*

We use MovieLens dataset to implement purposed recommendation system. MovieLens contain data about millions of users with their ratings to items. To test performance mean absolute error (MAE) was utilized.

For result and experiment, we have used 10M size dataset of MovieLens. It consist 10000054 ratings, 10681 movies rated by 71567 users. In this dataset three files are their movies.dat, ratings.dat, and tag.dat. ratings data file have three columns userId, MovieId, and tatings. User have only UserId for identification no other information are given. Each user at least 20 movies are rated.

- *Evolution Matrix*

To speed the performance of recommendation list it implemented on Hadoop. To improve speedup and efficiency of recommendation system MapReduce are utilized.

1. Accuracy Evolution

Three metrics are used to evaluate the accuracy, mean absolute error (MAE), mean average precision (MAP) and discounted cumulative gain (DCG). The lower the MAE presents the more accurate predictions, and the higher MAP or DCG presents the higher quality of the predicted Top-K service recommendation list.

Threshold value is calculated according to the no. of users like same item (movies). Accuracy in recommendation is increases when threshold decrease. The number of movies recommended by user is decreases when threshold value increases shows in fig 2.

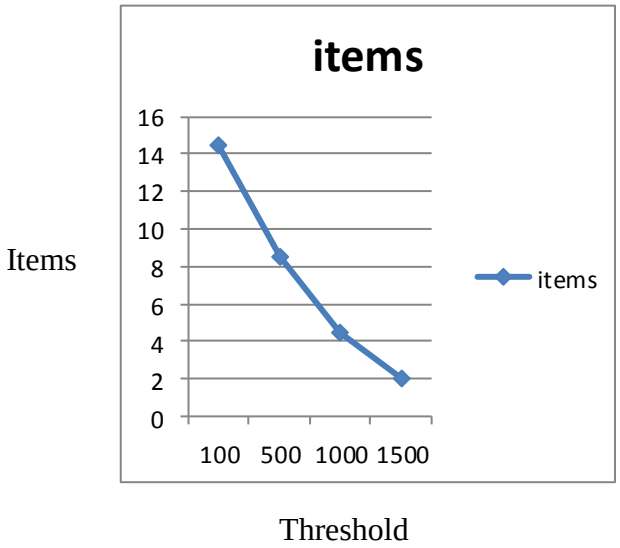

Fig. 2 Comparison graph based on Threshold value

2. Scalability Evolution

A well-accepted scalability metric, Speedup, is adopted to measure the performance in the scalability of recommendation

list. Speedup refers to how much a parallel algorithm is faster than a corresponding sequential algorithm, which can be defined as follows:

$$Speed\ up = \frac{T1}{Tp}$$

Where p is the number of processors, T1 is the sequential execution time. Tp is the parallel execution time with p processors. If the speedup has a linear relation with the numbers of nodes with the data size fixed, the algorithm will have good scalability.

The processor performance is measured in scalability which shows in Fig 3.

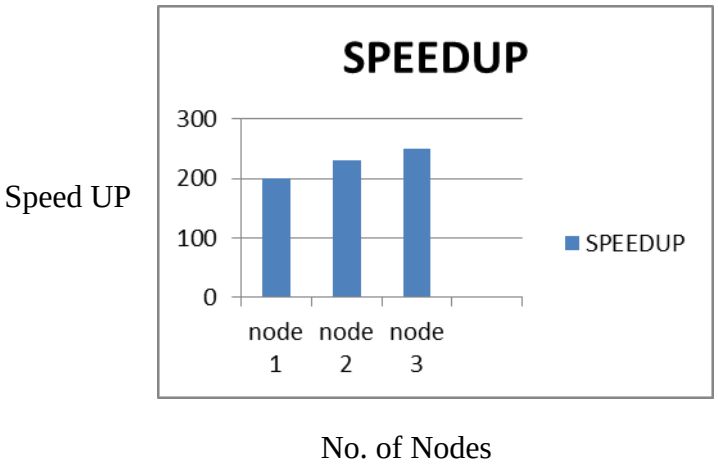

Fig-3: Speed Up with respect to no. of nodes

VI. Conclusion

In this paper, we have proposed a personalized user-based recommendation system for movies recommendation system. Our system aims at presenting a personalized service recommendation list and recommending the most appropriate service(s) to the users. Moreover, to improve the scalability and efficiency in "Big Data" environment, we have implemented it on a MapReduce framework in Hadoop platform.

With the development of cloud computing software tools such as Apache Hadoop, Map-Reduce, and Mahout, it becomes possible to design and implement scalable and efficient recommender systems in "Big Data" environment.

In our future work, we will do further research in how to deal with the case where term appears in different categories of a domain thesaurus from context and how to distinguish the positive and negative preferences of the users from their reviews to make the predictions more accurate.

The proposed system is more efficient in terms of complexity. And the system gives more accurate results or recommendations to the users.

VII. References

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, and Jinjun Chen, Senior Member, IEEE KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 12, DECEMBER 2014.

[2] X. Yang, Y. Guo, and Y. Liu, Bayesian-Inference Based Recommendation in Online Social Networks, IEEE TRANS. PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO.4, PP. 642-651, APR. 2013.

[3] 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.

[4] G. Adomavicius and Y. Kwon, New Recommendation Techniques for Multicriteria Rating Systems, IEEE Intelligent Systems, vol. 22, no. 3, pp. 48-55, May/June 2007.

[5] Z D Zhao and M. S. Shang, "User Based Collaborative Filtering Recommendation Algorithms on Hadoop," In the third Internation-alWorkshop on Knowledge Discovery and Data Mining, pp. 478-481, 2010.

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

[7] Magdalini Eirinaki, Malamati D. Louta, Member, IEEE, and Iraklis Varlamis, Member,IEEE "A Trust-Aware System for Personalized User Recommendations in Social Networks"IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 44, NO. 4, APRIL 2014 409.

[8] Sarwat, Justin J. Levandoski, Ahmed Eldawy, and Mohamed F. Mokbel IEEE LARS*: An Efficient and Scalable Location-Aware Recommender System Mohamed TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014.

[9] Y. Chen, A. Cheng, and W. Hsu, "Travel Recommendation by Mining People Attributes and Travel Group Types from Community-Contributed Photos," IEEE Trans. Multimedia, vol. 25, no. 6,pp. 1283-1295, Oct. 2013.

[10] C. Lynch, "Big Data: How do your data grow?" Nature, vol. 455, no. 7209, pp. 28-29, 2008.

[11] Y. Pan and L. Lee, "Performance Analysis for Lattice-Based Speech Indexing Approaches Using Words and Subword Units," IEEE Trans. Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1562-1574, Aug. 2010.

[12] G. Kang, J. Liu, M. Tang, X. Liu, and B. cao, "AWSR: Active Web Service Recommendation Based on Usage History," Proc. IEEE 19th Int'l Conf. Web Services (ICWS), pp. 186-193, 2012.

[13] G.M. Amdahl, "Validity of the Single-Processor Approach to Achieving Large Scale Computing Capabilities," Proc. Spring Joint Computer Conf., pp. 483-485, 1967.

[14] H. Liang, J. Hogan, and Y. Xu, "Parallel User Profiling Based on Folksonomy for Large Scaled Recommender Systems: An Implimentation of Cascading MapReduce," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 156-161, 2010.

[15] B. Issac and W.J. Jap, "Implementing Spam Detection Using Bayesian and Porter Stemmer Keyword Stripping Approaches," Proc. IEEE Region 10 Conf. (TENCON '09), pp. 1-5, 2009

[16] An Effective Recommendation Framework for Personal Learning Environments Using a Learner Preference Tree and a GA Mojtaba Salehi, Isa Nakhai Kamalabadi, and Mohammad B. Ghaznavi Ghoushchi, Member, IEEE IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES,VOL. 6, NO. 4, OCTOBER-DECEMBER, 2013.