VOL. 10, NO. 7, APRIL 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

OPTIMAL KEYWORD SEARCH FOR RECOMMENDER SYSTEM IN BIG DATA APPLICATION

J. Amaithi Singam and S. Srinivasan
Department of Computer Science and Engineering, Regional Centre of Anna University, Madurai, Tamil Nadu, India
E-Mail: amaithi24@gmail.com

ABSTRACT

Currently, online searching process increases and people searches new information in the search process. Most of the search engine gives additional supporting information. Recommender system involves in this process and implements as service. Service recommender system gives additional information to the user but if information grows then these process become a critical one. The proposed work analyses issues occurring when service recommender system implements in large data sets. This work proposes a keyword-Aware services Recommender method, to split the services to the users and mainly focused keywords from the user preferences. Hybrid Filter algorithm generates keyword recommenders from the previous user preferences. To implement effective results in big data environment, this method is implemented using the concept of Map Reduce parallel processing on Hadoop. Experimental results are shown the effective results on real-world datasets and reduce the processing time from large datasets.

Keywords: recommendation system, web search, big data, map reduce.

1. INTRODUCTION

A web search engine [9] is a software system that is designed to search for required information. The search results are generally presented in a line of results .The information may be a mix of web pages, images, document and other types of files. The input of the search engine may be a word, string or token. Data means collection of information, it may be human language or non-human language. This kind of information increases rapidly that cause's big data problem. Big data can be categorized into volume, varity, velocity. Volume means it can process terabytes to zettabytes data. Varity may be in the form of structure, unstructured, semi structure data. Velocity means streaming of data. The big data mainly faces two problems; they are scalability problem and accuracy problem. Scalability problem [22] occurs when the amount of processing time increases while processing of large amount of data. Currently we have enormous volume of structured and unstructured data, and while searching we could not get accurate results. For example, railway ticket reservation system contains terabytes of data. So that active user finds difficult to pick the correct data.

Recommendation system depicts additional information to the active user such as learning material Recommendation system [7], Recommendation system in real e-commerce, Music recommendation system in Apple iTunes, Amazon.com recommendation system, many more. The main role of the Recommendation system is information filtering. Recommendation system can be classified into two types, they are knowledge based recommendation and collaborative filtering [9] [10]. The widely used technique is collaborative filtering. The collaborative filter only focuses on the user preference of active user. Knowledge based recommendation system means asking a user about their own preference of the item and find out what active user required.

The recommendation system mainly uses four filters for information retrieval, 1. Demographic filter, 2.Content based filter, 3. Collaborative filter, 4. Hybrids filter

Demographic filter provides the additional information to the required item. The widely used another filter is content based filter [21]. The main work of the content based filter is item oriented similarity, it provides widely desired item in the past. Collaborative filter (CF): is most effective filter in recommendation system. The collaborative filter depends on the user preference, user interest or user taste recommendation. The collaborative filter automatically appends the up-to-date information. Compared with content based filter it only concentrates on the item but does not focus on user preference. CF [14][17] can be classified into two types, they are memory based and modern based. Memory based is based on the database, also used item to item or user to user based filtering techniques it fully depend on the correlation based techniques to predict the future item. A model based technique does not use whole data set to predict the item. The most widely used techniques in model based are clustering based techniques and single value decomposition techniques.

Our proposed work is optimal keyword search for recommendation system in big data application. Here we use Hybrid filter. Hybrid filter is a combination of both content and CF techniques. Here we consider Hotel reservation system that provides best hotel in preferred area based on the user request. Finally we compare the efficiency and scalability [23] for different datasets.

2. RELATED WORK

The service recommendation [1] system provides additional information to the user. Currently the amount of data has increased rapidly that yields big data problem. In traditional service method faces two problems when processing large amount of data. They are efficiency of

data and scalability problem. Moreover most of the traditional recommended systems focus on the ranking and ratings of service with different services and different user. Existing system does not consider the user preference [2], only focuses on the item of the service. Shunmei Meng et.al. (2013) proposed the service recommendation system [5] used the collaborative filter algorithm to generate the approximate recommendation to the active user. That improves the scalability and efficiency of the data when processing large amount of data. Works are implemented on Hadoop using MapReduce [12] framework with real time dataset. Finally results are compared to the existing system result of the recommendation system. And to provides accurate data and scalability of large amount of data.

The recommendation system predicts [6] the future user taste based on the active user preference. Rapid growth of data the user cannot pick out the required data in internet. Reema Sikka *et al.* 2012 discussed about the filtering approaches and different types of recommendation system. To improve the accuracy of the data and also used the collaborative filter to predict the user taste to generate the recommendation to the active user. And the item based approaches to identify the relation between the past and present user preference, also implemented the user based approach to predict [19] the user taste. Finally, to compare the different algorithms such as Random prediction, Frequent Sequence, Collaborative Filter, and Content based filter. The collaborative filter algorithm was used. The collaborative filter algorithm is to produce efficient data and improve the scalability problem.

The recommendation system e-learning [3] is based on the user data and evaluation of result. Currently, Resources are available bulk of learning material in online or offline. The peoples are selecting the correct required material in internet. Also study about various recommendation techniques to explain the four filters, they are Demographic filter, Content based filter, Collaborative filter and Hybrids filter. Rubina Parveen *et al.* 2012 discussed the drawbacks and advantages of the entire filter.

The viewer's choices on the product advertisement are important part of the market. This part generates the recommendation [11] [13] to the viewers. Atisha Sachan *et al.* 2012 is proposed the two data collection methods. They are implicit and explicit data collection. Also the filter can be classified into four types, they are 1.Demographic filter, 2.Content based filter, 3. Collaborative filter, 4. Hybrids filter. And discuss about the four filters but mainly focused on the more powerful and effective collaborative filter. The major challenges of this filter are 1.Cold start problem, 2.Data Sparsity, 3.Scalability, 4.Accuracy of data problem.

A recommendation system in an e-learning context [4] is efforts to intelligently recommend to the active user (learner) based on the action of the previous user (learner). Also this recommendation system is based on the online activity such as reading posted message on internet. These recommendation systems have been tried in an e-commerce [13] to purchase the quality goods. Dhoha Almazro *et al* (2013) addresses the use of web mining to build the online activities based on user (learner) access History to improve course material navigation as well as improve the processing time. Also e-learning compared to the traditional system of face-to-face style teaching and learning. It provides more benefits than the face-to-face learning. Here the content based filter is used. The content based filter to use the algorithm of vector space model for similarity computation to predict the similar item based on the user rated item that improve the accuracy of data.

The information about the product is increasing with exponential rate in e-commerce industry [15]. This environment has complexity to find the optimal information in majority of data collection. The customer can get benefit by receiving optimal information about the products which they are customer likely to buy. And discussed about the different type of recommendation system filters [16] method and mainly focused on the algorithm of collaborative filter and comparing each and every approaches of collaborative filter. They are 1.Active filter, 2.Passive filter, 3.Item based filter. The Active filter uses the peer to peer approaches. Passive filter focuses on the implicit collection of information. The item based filter only addresses the item of the user preference. Jong Seo Lee (2012) describes the different kind of recommendation system. They are 1.Musical recommendation system in apple iTunes, 2.Recommendation system in amazon, 3.Recommndation system in like-i-like, 4.Music Surfer beyond Existing Music Recommendation system. And also describe the following approaches, they are 1. Clustering model, 2.Bayesian network model, 3.regression approaches.

Bigtable [18] stored the structure data in distributed storage system manner to scale the very large size. Currently, it can store beyond the terabytes of the data. Google store, Web indexing, Google earth, Google Finance also used a BigTable for storing structured data. The Bigtable has successfully provided a flexible, high performance solution for the entire Google product. F. Chang *et al.* 2008 described the data model such as sparse, distributed and persistent multidimensional data. All data models are stored in rows and column manner. And used analogous of three level hierarchies and evaluates the performance of data storing and recovery compared to the traditional methods.

3. ARCHITECTURE

We gathered the existing user's reviews and stored it into the dataset in Figure-1. The Data set was downloaded from UCI repository. It is named as Hotel Reservation System. The attributes of this dataset are Hotel name, Date of the Review and Users review. The preprocessing step is used to remove the noise from dataset. After removing the noise the data is stored into database.

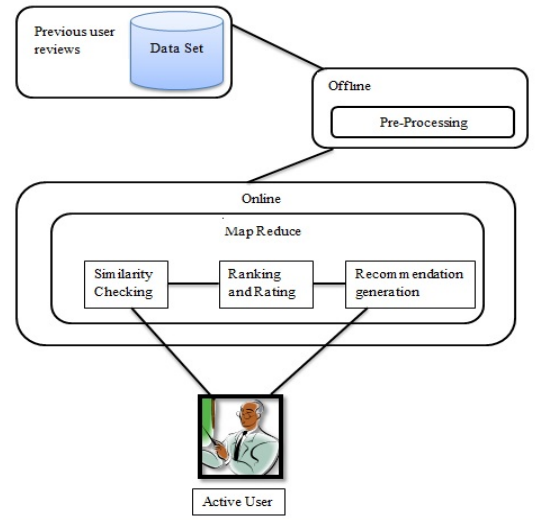

Figure-1.

These processes are under the offline. Then the Active user gives their preferred item and rated of the item. That preferred item was compared to the previous user preference. If the Active user preferred item is similar to the previous user item then the required similar item is rated [20] and ranked. Finally we generate the Recommendation to the Active user. Online environment similarity checking, Ranking and Rating, Recommendation Generation [8] are performed using map reduce concept.

Preprocessing

The previous user's reviews are stored into the Dataset. The Dataset was downloaded from UCI repository. The size of the Dataset is 257MB. The Data set was stored in the text files. Main attributes of the Data set is Hotel Name, Year, Previous user Reviews.

The main function of preprocessing is to remove the noise from the previous user reviews and required keywords are extracted from the data set using porter stemmer algorithm.

Porter stemmer algorithm

The porter stemmer algorithm is used to remove the suffix word from the previous users review. And the common morphological also elite that term called stem. The document represents the term or vector form. Here the previous user review keywords are classified using porter stemmer algorithm.

Previous user review (PUR) keyword1= neatness
Previous user review (PUR) Keyword 2=neat

Step-1: compare {(PURK1), (PURK2)}
Step-2: REMOVE suffix→ ness
Step-3: Stem=neat

Similarity checking

The previous user reviews are extracted from Dataset and stored the database. The attributes are Hotel name, Date, keywords, count. The active user gives their required Keywords with rating. The previous user's keywords (PUK) are compared to the Active user's keywords (AUK). If the keywords are exactly same then the keywords are stored in the matrix using Pearson correlation coefficient algorithm this method is called exact similarity Otherwise the method is said to be approximate similarity.

Approximate similarity method

Set theorem is used to find out the similarity between the active user keywords and the previous user keyword.

$$\mathbf{Sim\ (AUK, PUK)} = \frac{|\mathbf{AUK \cap PUK}|}{|\mathbf{AUK \cup PUK}|} \longrightarrow \quad (1)$$

Exactly similarity (Collaborative filter)

The required keywords of the active user and previous user keywords are transformed into the form of n dimensional weight vector. And Similarity and weight was calculated using Pearson correlation coefficient algorithm.

Pearson correlation coefficient algorithm

Correlation between the set of keywords is a measure of how well they are related to the similar. Pearson Correlation Coefficient Algorithm is used to measures the linear relationship between two sets of Keywords and also measures the user preference.

$$\mathbf{Sim\ (a,p)} = \frac{\Sigma_{k\in K}(r_{a,k} - \bar{r_a})(r_{p,k} - \bar{r_p})}{\sqrt{\Sigma_{k\in K}(r_{a,k} - \bar{r_a})^2}\sqrt{\Sigma_{k\in K}(r_{p,k} - \bar{r_p})^2}} \longrightarrow \quad (2)$$

where K is the set of Keywords rated by both users, $r_{a,k}$ and $r_{p,k}$. r_a and r_p denoted the rating of both active users end previous users. r_a and r_p represents the mean value of active user and previous user. This is under the Collaborative filter.

$$\mathbf{W\ (k1,k2)} = \frac{\Sigma_{u\in U}(r_{u,k1} - \bar{r_{k1}})(r_{u,k2} - \bar{r_2})}{\sqrt{\Sigma_{u\in U}(r_{u,k1} - \bar{r_{k1}})^2}\sqrt{\Sigma_{u\in K}(r_{u,k1} - \bar{r_{k2}})^2}} \longrightarrow \quad (3)$$

where k1, k2 keywords are selected by user's u, here U denotes the set of Users.r_{k1} and r_{k2} represents the mean value of the keywords k1 and k2.

Rating and ranking (content based filter)

The rating value can be calculated by the **si**milar users. Then the service recommendation list will be generated from that user can collect the highest rating service using nearest neighbor's algorithm.

Top k with nearest neighbor's algorithm

$$\mathbf{G_{a,k}} = \frac{r_a + \Sigma_{u\in K}(r_{u,k} - r_u) \times W(k1,k2)}{} \longrightarrow \quad (4)$$

$\Sigma_{u \epsilon K} W(k1,k2)$

Where Ga,k is the prediction for the active user a for keyword k, W(k1,k2) is the similarity between users a and p, and K is the neighborhood or set of most similar users.

Map reduce

Map Reduce [1] [3] [23] is used to execute the data in parallel manner. Here the Similarity checking, weight generation, rating and ranking are executed in parallel [20].

4. EXPERIMENTAL RESULTS

This section describes the performance analysis to validate the proposed algorithm.

Accuracy

Experimental results demonstrate the efficiency of the proposed Pearson correlation algorithm. Figure-2 shows the comparison of cosine theorem and the Pearson correlation coefficient algorithm.

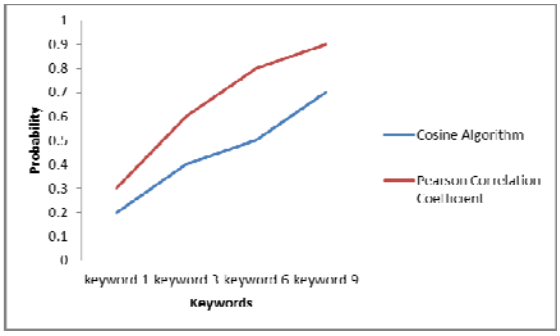

Figure-2. Algorithm comparison.

Figure-2 depicts the comparison of Cosine and Pearson correlation. For keyword 3 our proposed algorithm shows a probability of 0.6 whereas cosine based algorithm shows the probability of 0.4. Our proposed algorithm increases the probability 0.2 when we compared to cosine based algorithm.

Scalability

When the data set is increased, we have scalability problem.

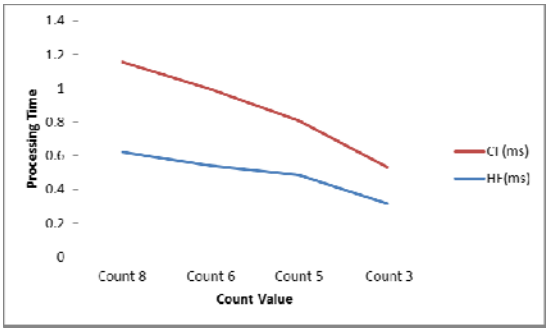

Figure-3. Scalability checking.

Figure-3 depicts the comparison of Collaborative filter and hybrid filter. For Count 3 our proposed filter shows a processing time of 0.2141ms whereas Collaborative filter shows processing time of 0.3172ms. Our proposed filter reduces the processing time upto 0.1031ms when compared to Collaborative filter.

5. CONCLUSIONS AND FUTURE ENHANCEMENT

We have proposed a Service Recommendation System on Map Reduce concept. In this work keywords are used to indicate users'(Active and existing) preferences, and a hybrid Filtering algorithm is adopted to generate appropriate recommendations. The active user gives the preferences by selecting the keywords with rating from the keyword candidate list. The preferences of the previous users can be extracted from their reviews for services corresponding to the keyword-candidate list. Our method aims to improve the scalability and efficiency of Big Data environment, Implemented on map reduce concept.

REFERENCES

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen and Senior Member. 2013. "KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Application", IEEE Transaction on Parallel and Distribution System.

[2] X. Yang, Y. Guo and Y. Liu. 2013. "Bayesian-inference based recommenda- tion in online social networks," IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 4, pp. 642-651.

[3] Z. Zheng, X.Wu, Y.Zhang, M.Lyu and J.Wang. 2013. "QoS Ranking Prediction for Cloud Services," IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 6, pp. 1213-1222.

[4] Dhoha Almazro, Ghadeer Shahatah and Lamia albdulkarim. 2013. "A Survey Paper on Recommendation System".

[5] M.Alduan, F.Alvarez, J.Menendez and O. Baez. 2013. "Recommender System for Sport Videos Based on User Audiovisual Consumption," IEEE Transactions on Multimedia, Vol. 14, No.6, pp. 1546-1557.

[6] Reema Sikka, Amita Dhankhar and Chaavi Rana. 2012. "A Survey Paper on E-Learning Recommendation System",International Journal of computer Application (0975-888). Vol. 47, No.9. June.

[7] Rubina Parveen, Anant Kr. Jaiswal and Vibhor Kant. 2012."E-Learning Recommendation System- A Survey", International Journal of Engineering

Research and Development, Vol. 4, Issue 12 November.

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

[9] L. Wu, X. Xiao, D. Deng, G. Cong, A. D. Zhu and S. Zhou. 2012. "Shortest Path and Distance Queries on Road Networks: An Experimental Evaluation," Proc.VLDB Endowment, vol.5, no.5, pp.406-417.

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

[11] Y. Chen, A. Cheng and W. Hsu. 2012. "Travel Recommendation by Min- ing People Attributes and Travel Group Types From Community- Contributed Photos", IEEE Transactions on Multimedia, Vol. 25, No.6, pp. 1283-1295.

[12] Atisha Sachan and Vineet Richariya. 2012. "A Survey on Recommender System based on Collaborative Filtering Technique", International Journal of Innovation in Engineering and Technology(IJIET).

[13] Y. Pan and L. Lee. 2010. "Performance analysis for lattice-based speech indexing approaches using words and sub word units," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 6, pp. 1562-1574, 2010.

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

[15] H. Liang, J. Hogan and Y. Xu. 2010. "Parallel User Profiling Based on Folksonomy for Large Scaled Recommender Systems: An Impli mentation of Cascading MapReduce," In: Proceedings of the IEEE International Conference on Data Mining Workshops, pp. 156-161.

[16] Jong Seo Lee. 2011. "Survey of Recommendation System (Collaborative filter)", California Polytechnic State University.

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

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

[19] F. Chang, J. Dean, S. Ghemawat and W. C. Hsieh. 2008. "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems, Vol. 26, No. 2(4).

[20] Y. Zhu and Y. Hu. 2006. "Enhancing search performance on Gnutella-like P2P systems," IEEE Transactions on Parallel and Distributed Systems, Vol. 17, No. 12, pp. 1482-1495.

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

[22] G. Linden, B. Smith, and J. York. 2003. "Amazon.com Recommendations: Item-to-Item Collaborative Filtering," IEEE Internet Computing, Vol. 7, No.1, pp. 76-80.

[23] G. M. Amdahl. 1997. "Validity of the single-processor approach to achieving large scale computing capabilities", Proceedings of spring joint computer conference, pp. 483-485.

[24] W. Hill, L. Stead, M. Rosenstein and G. Furnas. 1995. "Recommending and Evaluating Choices in a Virtual Community of Use," In CHI '95 Proceedings of the SIGCHI Conference on Human Factors in Computing System, pp. 194-201.

[25] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom and J. Riedl. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," In: CSCW '94 Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175-186.