Volume-5, Issue-1, January -2016 • ISSN No 2277 - 8160

Research Paper Engineering

# Hadoop Based Collaborative Recommendation System

Harsh S. Varudkar Student, Shah & anchor kutchhi engg. College, Mumbai

ABSTRACT *21st century is an age of internet and information. Growth of internet in world increases the traffic almost 3 times more. Ecommerce market, social media and online educations are some field which leads rapid growth of online presence in past few years. These lead huge amounts of information growth also Users' tendency towards market is changing rapidly. Hence the competitions between industries are on a peak in current situation. Hence Customer has been induced. It is difficult to find a desired interest on internet by just surfing. Information filtering system is a system that removes dispensable or unnecessary information before presenting it to the human user. It is a way of sifting through the overabundance of data on the web. These information filtering system have a subclass called recommender systems or recommendation systems that predict the 'rating' or 'preference' that user would give to an item*

KEYWORDS : Recommendation System, BIG Data, Collaborative filtering, Pearson correlation

## INTRODUCTION

Recommender system can be defined as the derive class of information filtering system which attempts to give the guidance to the users regarding the useful services based on their personalized preferences, past behaviour or based on their similar likings with other users. It is becoming difficult to capture, store, manage and analyze such big data that affects the service recommender systems with issues like scalability and inefficiency. Also many existing service recommender system provides the same recommendations to different users based on ratings and rankings only, without considering the taste and preference of an individual user.

Information filtering system is a system that removes dispensable or unnecessary information before presenting it to the human user. It is a way of sifting through the overabundance of data on the web. These information filtering system have a subclass called recommender systems or recommendation systems that predict the 'rating' or 'preference' that user would give to an item. Service recommender systems provide appropriate recommendations of services to the users. These service recommender systems have become popular in variety of practical applications like recommending the users about movies, contents, articles, books, books, search queries, social information and items.

In user to user Collaborative recommendation system users' previous behavior is used for finding similarity matrix with respect to other users, in such case if we have large number of user then it will preferred to use item to item similarity to generate recommendation [1-3], in user to user similarity users' interest has been predicted from its past behavior, and according to other similar users behavior has been analyzed, users' profile and its past logs are taken for analyzing similarity[4]. Item to item recommendation system is basically more effective in such scenario where we have huge numbers of users as well items, here rather than using users' profile, item similarity has been analyzed. Item type its logs, popularity and description are the main pillars for analyzing similarity [5]. But cold start problem effects in this scenario, the newly arrived item has been lacked exposer here heuristic methods has been used to generate recommendation [6]. Unique classifier has been generated by providing more information.

## REVIEW OF LITERATURE

Recommendation systems algorithms has three categories, Memory based, Model based and Hybrid. Traditional recommendation systems has lack of scalability as it is not capable to handle large scale data. User diversity is also one of the important factor for understanding user interest which is also ignored [8]. The keyword used to show users' preference and collaborative filtering algorithm was adopted to generate appropriate recommendation. Keyword-Aware Service Recommendation method to address challenges of different users' and their diverse preference evaluation. The results demonstrated that Keyword-Aware Service Recommendation method significantly improved recommendation system with respect to current traditional recommendation approaches. [4]

A collaborative filtering recommendation method name TyCo. It features of finding neighbor of users based on user degrees in user groups instead of co-rated items or similar users of items. The idea behind the method is of object typicality from cognitive psychology [3]. Popular items may not be users' requirement. They argued suggested that the model should be independent from domain in proposed e-learning recommendation model [10]. A dataset has few items and lots more users then item based model will be good choice and in vice –versa scenario user base model will be better.

For better efficiency and performance in practice there are multiple model used for computation linearly [12]. In distributive environment on cloud platforms are one of good platform to manage huge amount of data in recommendation system and to achieve the scalability in service recommendation system based on keyword aware and also to solve inefficiency in handling large amount of services during recommendations. Also development of cloud computing software tools such as apache Hadoop, map reduce and mahout made possible to design and implement scalable recommendation system in big-data environment. A problem faced by recommendation system is its scalability when the dataset volume is very much high than computation cost is also increase. In order to solve scalability issue collaborative filtering can be implemented on cloud platform, HADOOP which solve problem for large scale of data grows [13]. In User-based and item-based filtering, the final prediction has been resulted by from 3 sources, 1) prediction made by same user on different item 2) predication by other user of same item 3) data predication based on other users' with similarity of rating on similar items. In this paper we focused on 3 primary research approach. 1) Finding similarity by pearson correlation technique. 2) Analysis of user-to-user and item-to-item collaborative recommendation system. 3) Implementing such system on hadoop based cloud environment.

## PROPOSED WORK

Here we are going to propose a scalable and robust recommendation system, this work has 3 stage of implementation 1) Gather users' rating on item by implicit and explicit way of data gathering. Explicit data gathering contain the dataset of users' given rating and feedback about the items. Implicit data gathering involves user behaviour data i.e. clickstream, session, bounce rate study. 2) As per given data accumulated user neighbourhood has been determined from these the user similarity has been determined. User similarity over a particular item will be analysed by Pearson correlation method. This technique is very popular and easy to implement. 3) 3rd step is to give recommendation. There are 2 ways to give recommendation, to anonymous user by taking their interest and recommending items, to known user by finding the similarity taste of user based on previous item rating. There are many technique to find similarity like Manhattan distance measure, Euclidean. Normally user interest data have many grade inflation data on which the higher values will definitely dominate the calculation hence such technique is not capable to produce the robust recommendation. In 1st stage users' rating has been calculated from explicitly and implicit way. For Explicit users we use dataset which contain users' previous rating are preference. For implicit data collection user behaviour data, clickstream data is been collected and converted into form which is supported to recommendation system. We use apache flume for implicit data collection.

Apache flume is used to gather bulk data and store in hdfs directly the data collected is in raw format hence data has to be converted into form of rdbms version hence apache hive is used to portioned the raw bulk data from hdfs and make it in tabular format. Apache hive distinguish data in related format and further we can also export the files into .csv file also as input in recommendation system.

2nd stage of proposed system is to find the similarity between users by using Pearson correlation technique. Pearson correlation technique is very effective in case of grade inflation.

**Stage 1, 2 and 3 has been described in fig. 1.**

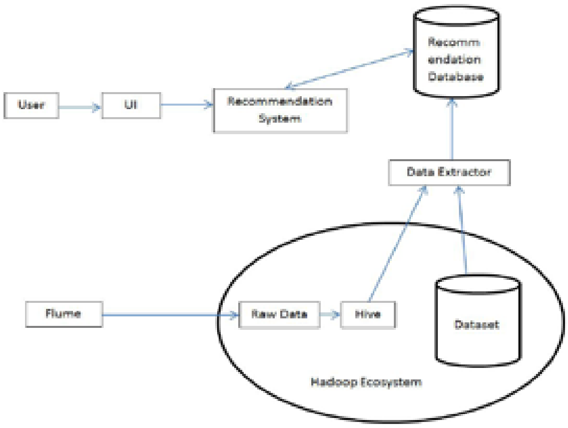

**Fig.1 Hadoop based Collaborative Recommendation system**

The exported CSV files contains the users' rating in user_id, item_id and rating format. The output has been stored on different file. Mapper read the input csv file and reducer produce the rating. All reducer calculate the similarity and write into output file. The formula for the Pearson Correlation Coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Pearson correlation technique is one of very popular technique to find similarity. Pearson algorithm is highly recommended in grade inflation scenarios. Pearson algorithm cannot be used if there is a sparse data. Hence the in recommendation system only data has been taken which has dense data and avoid blank fields. Formula for Pearson is as follows, where Iu is the items rated by u:

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v}(r_{ui}-\mu_u)(r_{vi}-\mu_v)}{\sqrt{\sum_{i \in I_u \cap I_v}(r_{ui}-\mu_u)^2}\sqrt{\sum_{i \in I_u \cap I_v}(r_{vi}-\mu_v)^2}}$$

µu should be computed just over the ratings in Iu∩Iv, Pearson technique produced similarity is between 0 & 1. If the result is near 0 then the users are not similar but is results comes neat 1 then the users are similar and neighbor to each other. 1 depict perfect similarity between users.

In 3rd stage the recommendation has been generated and presented to users, For anonymous user we will take interest based data and based on that the similar recommendation has been generated we can also use user behavior i.e. clickstream data to understand the user interest.. For known user or registered user we will use previous interest based data and according to generate the recommendation. The result can be show on webpage and also stored on different file.

## CONCLUSION & FUTURE WORK

This paper focused on collaborative recommendation based on hadoop. The hadoop components like apache flume and apache hive are very useful for collection of data and understanding the behavior of user. The map-reduce will also use parallel computation which will make this recommendation system more robust and scalable. The mahout library is been used for implementing the mahout over hadoop system for better performance.

Only rating will not depict the user taste over many product users behavior like user visits over various pages, clickstream data also make good contribution to generate recommendation. For item-based for anonymous and registered/known user, different item-item matrix may formed to find the similarity between users. Recommendation system used here implemented Pearson correlation technique to find similarity these may further exceed by using the other technique as per scenario proposed. The proposed system use collaborative recommendation system which may further exceed by proposing hybrid recommendation system in case of multiple feature of different recommendation system required to implement.

## REFERENCES

[1] Bellogin, A., Castells, P., & Cantador, I.." Neighbor selection and weighting in user-based collaborative filtering": A performance prediction approach. ACM Transactions on the Web (TWEB), 8(2), 12. [2] Shi, Y., Larson, M., & Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Computing Surveys (CSUR), 47(1), 2014, pp 3-10. [3] Ghuli, P. ; Dept. of Comput. Sci. & Eng., R.V. Coll. of Eng., Bangalore, India ; Ghosh, A. ; Shettar, R., "A collaborative filteringrecommendation engine in a distributed environment" [4] Jinjun Chen, Wanchun Dou, Xuyun Zhang, "KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Application," IEEE Transactions on Parallel and Distributed Systems, [5] Pazzani, Michael J. "A framework for collaborative, content-based and demographic filtering." Artificial Intelligence Review 13, [6] Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. Machine learning, 27(3), 313-331. [7] Xiaoyuan Su and Taghi M. Khoshgoftaar," A Survey of Collaborative Filtering Techniques" Hindawi Publishing Corporation Advances in Artificial Intelligence Volume 2009 [8] Jun Wang, Arjen P. de Vries, Marcel J.T. Reinders "Unifying User-based and Item-based Collaborative Filtering Approaches by similarity fusion" SIGIR '06 [9] Xiwei Wang, Erik von der Osten, Xuzi Zhou, Hui Lin,"A Case Study of Recommendation Algorithms", 2011 International Conference on Computational and Information Sciences [10] Saman Shishehchi, Seyed Yashar Banihashem, Nor Azan Mat Zin, Shahrul Azman Mohd. Noah, "Review of personalized recommendation techniques for learners in e-learning systems [11] Paritosh Nagarnaik, Prof. A.Thomas," Survey on Recommendation System Methods", IEEE SPONSORED 2ND INTERNATIONAL CONFERENCE ON ELETRONICS AND COMMUNICATION SYSTEM (ICECS 2015) [12] Neethu Raj,Suja Rani M S, "An Overview of Content Recommendation Methods"International Journal of Innovative Research in Computer and Communication Engineering. [13] Ruchita V. Tatiya, Prof. Archana S. Vaidya . "A Survey of Recommendation Algorithms" IOSR Journal of Computer Engineering [14] Robin Burke, 1. Department of Information Systems and Decision Sciences, California State University, Fullerton, CA, Hybrid Recommender Systems: Survey and Experiments, User Modeling and User-Adapted Interaction, November 2002, Volume 12, Issue 4, pp 331-370, http://link.springer.com/article/10.1023/A:1021240730564 [15] Saman Shishehchi, Seyed Yashar Banihashem, Nor Azan Mat Zin, Shahrul Azman Mohd. Noah, "Review of Personalized Recommendation Techniques for Learners in E-learning Systems"