Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-10

A Survey on Recommendation Algorithm for Movie Recommendation on Cloud

[Swati Pandey, Dr. T. Senthil Kumar]

***Abstract*— In current era, Web is the best source for getting any information or making decision on something. People get online suggestions before making any decision such as buying any product, booking movie tickets etc. In such cases recommendation systems play important role. Recommendation System works on the data about users and items which has to be recommended. Due to huge size of data, distributed systems come into existence.**

***Keywords*— Recommendation System, Hadoop, HBase, Mahout, Map-Reduce**

I. Introduction

Now a day, for getting solutions of problems people prefer web. Search engines fulfill their requirements partially. Search engine does not give result according to users preference, users taste or content of item required by the user. It provides all possible outcomes related to users query. Hence for making an efficient decision regarding an item or person, Recommendation systems are needed. Recommendation systems are used for the personalization of information. It helps users for making decisions regarding an item or a person [12]. For example which movie you should watch which book you should buy etc. There are websites through which users can give their ratings or views about a particular item or anything. Input for recommendation system can be [11]:

- Rating, this is the opinion of user for an item. It can be collected implicitly or explicitly.
- Demographic data, which is information about user such as age, gender, occupation. Normally is collected explicitly.

Output of the recommendation system can be [11]:

- Prediction: It represents predicted opinion for user for an item. It should be in same scale as of input rating.
- Recommendation: It includes list of top N homogeneous items recommended for an active user.

For recommendation purpose we need to collect data about related items or things, Users or user groups and their views or ratings. Recommendation system uses Machine learning algorithms for recommending items/person. Since dataset for recommendation purpose, may be large hence executing that dataset on a single node is not an efficient way. For optimizing execution and getting solutions fast we go for cloud. Cloud is a specialized form of distributed system. Distributed system consists of a collection of autonomous computers connected through a network and distributed middleware, which enables computers to coordinate their activities and to share the resources of the system, so that user perceives the system as a single, integrated computing facility [1]. Hadoop is an open source cluster based frame work which is used for writing and running distributed application that process large amount of data [3]. It is a framework for Map-Reduce programming. Map-Reduce programming model is used to process and generate a large dataset according to Map and Reduce function [2]. Whole dataset is divided into key/value pairs. Map function has been specified by the user, which processes key/value pairs to generate intermediate key/value pairs. After processing, Reduce function merges all intermediate results associated with the same intermediate key. Then final result has been given to the user [2].

Rest of the paper has been arranged in different sections. Section 2 describes the basic procedure for building a Recommendation system. Section 3 elaborates the related work which has been done. Section 4 puts focus on the outcome of the survey. Section 5 describes about the proposed customization of movie Recommendation system.

II. Procedure for building Recommendation System

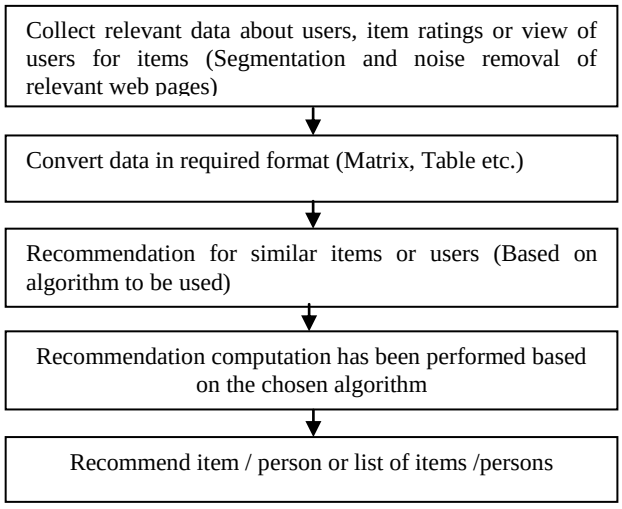

Figure 1: Basic Procedure for Recommendation System

Swati Pandey
Amrita University, Coimbatore -461112, Tamilnadu, India

Dr. T. Senthil Kumar
Amrita University, Coimbatore -461112, Tamilnadu, India

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-10

For recommendation engine first we need to collect data. Data can be collected implicitly or explicitly. That data may contain unwanted information or data also, hence to remove that we can use noise removal techniques. After getting relevant data, we convert that in our required form (generally matrix, tables). Now data is ready to use. With the help of an algorithm, we can recommend item/person or predict which item has to be liked by required user.

Basic procedure for building a recommendation system can be described as figure 1.

III. Related Work

A. *Segmentation and Noise Removal*

To extract relevant information through web page, Segmentation plays an important role. For movie Recommendation System, implicit or explicit ratings of users on movie are required. Sometimes user writes comments on movies. Hence to get rating from their comments we need to extract their comment from particular web page. A web page can contain multiple parts with different information content [4]. Hence page has to be segmented and non-relevant parts have to be removed for getting precise results. For recommending goods, we need to get ratings or views of different users about items; we need to find similar users. For collecting this data only relevant part of webpage has been accessed. Some research work that has been done for segmentation and noise removal is shown in table 1.

B. *Machine Learning Approaches for movie Recommendation System*

There are different machine learning approaches for recommendation purpose. These approaches can be categorized as following [8]:

- Random Prediction Algorithm: According to this approach, an item has been taken randomly form the large set of items and recommends to the user. Its accuracy depends on the luck. Hence this algorithm is failure.

- Frequent Sequence Algorithm: It recommends the item to user based on the past rating or views of user for items. Its accuracy depends on the users past ratings.

- Content Based Filtering: It focuses on the content and attributes of items. Item whose content correlates the most with the content of item that has already viewed in past and which satisfies user preference has been recommended.

- Collaborative Filtering: This algorithm identifies users that have relevant interest and preferences by calculating similarities and dissimilarities between users profile.

- Hybrid Approach: It combines both Content based as well as collaborative filtering approach.

Collaborative filtering can be categorized into two User based collaborative filtering, Item based collaborative filtering. User based collaborative filtering is also known as nearest neighbor collaborative filtering. In this, statistical techniques are used to find nearest neighbor of target user. These systems use algorithms to calculate the similarity between user profiles of nearest neighbors to produce prediction whether target user will like a particular item or to recommend top-N items to target user [9]. User based similarities calculations are performed using row-wise. Item based collaborative filtering focuses on the neighborhood of the similar items. Basic idea is that mostly user purchases items that are similar to the one s/he already bought in the past [10]. In this approach, we can make criteria for similarity metric, for example item rating can be a metric. If we calculate similarity based on content of the item then it becomes content based. Some methods and survey which have been performed on Movie dataset has been given in table 2.

IV. Discussion

With the help of work that has been done, it can be stated that, content based filtering techniques are good for the data having content as main criteria and recommendation has to perform on same type of items which have been previously bought by target user. Collaborative filtering is good when we have sufficient number of users whose preferences are almost similar to each other, sufficient number of items which are rated by user. Basically if item rating is a metric for recommendation and prediction then collaborative filtering performs well. User based collaborative filtering has limitations related to scalability. Compare to user based filtering, item based algorithm sparse better and scale well. But its shortcoming is the cost to build item-item matrix. For similarity evaluation, computer takes more computing time and resources.

When data set is large then process may takes hours for computations. This problem can be solved by cloud platform, eg. Hadoop. Even though Hadoop can handle large amount of data efficiently, but results may not be accurate. This is overcome by combining the user-based recommendation results and the item-based recommendations. But each phasing its own disadvantages, So by applying the clustering technique in the combined result we can get the accurate recommendation.

V. Conclusion

This paper proposes a Movie recommendation system using combined Collaborative filtering algorithm (User based collaborative filtering using Pearson correlation similarity as well as item based collaborative filtering using on open source

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-10

cloud environment Hadoop using Mahout library and HBASE as database. Hadoop can evaluate and generate large amount of data. Mahout [13] is machine learning library which supports collaborative filtering well. . Hadoop uses HDFS file system, which stores data as flat files. HDFS follows write once read many ideology, it does not support random read write problem. To overcome the problems of HDFS, we are using column based, NOSQL database HBASE which stores data in key value pair. It provides low latency access to small amount of data within a large data set. Hence system will provide more accurate result for large Movie data set also.

TABLE 1: RELATED WORK IN NOISE AND SEGMENTATION

Author	Method/ Algorithm	Merit	De-merit
Christian Kohlschütter , Wolfgang Nejdl [13]	Densitometric approach for webpage segmentation which based on token density in a text fragment	Focus on low level properties of the text, Detects duplicate blocks and Non-Duplicate blocks also	Accuracy for detecting duplicate blocks is 61.7%
David Fernandes, Edleno S. de Moura, Altigran S. da Silva, Berthier Ribeiro-Neto, Edisson Braga [6]	Method aligns the DOM trees of Web pages of a site in Order to uncover their implicit structure.	Useful to segment web sites which are data-intensive. Segments the web pages and also able to cluster the segments into classes.	Less efficient for generating rules that can be used for string manipulation.
Jan Zeleny, Radek Burget [7]	Method combines vision based segmentation and template based clustering algorithm	Precise, Fast	Accuracy
Erdinç Uzan, Hayri Volkan Agun, Tarik Terlikaya [14]	Hybrid approach for extracting informative content which contain 2 steps: 1. Discover informative content using decision tree learning 2. Extract rules obtains from the (1)	Effective for structured as well as semi structured data	First stage of this approach is appropriate where time performance is not important.
Fei Hu, Ming Li, Yi Nan Zhang, Tao Peng, Yang Lei [15]	Reduces noise in web pages based on word density	Works well with pages that do not meet the XML specification, Less time consuming	When the topic content has small amount of words, the purification is not ideal.
Zhao Cheng-Li, Yi Dong-Yun [16]	Eliminate noises by Style Tree Model (DOM tree based)	Adaptive, Fast	Less efficient for generating rules that can be used for string manipulation.
Shekhar Babu Boddu [17]	Determine spatial locality (vision based)	Efficient	Less efficient for generating rules that can be used for string manipulation
Hiroyuki Sano, Shun Shiramatsu, Tadachika Ozono [18]	Method is comprised of 3 steps: 1. Layout template detection 2. Division into minimum blocks and detecting title blocks 3. Combination into web content bits	Suitable for extracting title blocks for segmentation	Effective when high precision of title block and high re-call of deciding non-title block
Renato Dominguez Garcia et al. [19]	Approach improves topic Exploration in blogosphere by detecting relevant segments.	Efficient for getting relevant data	Automatic Segmentation tool which has been used is not perfect.

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-10

TABLE 2: RELATED WORK IN MOVIE RECOMMENDATION ALGORITHM

Author	**Method**	**Merit**	**Demerit**	**Comment**
Zhi-Dan Zhao, Ming-Sheng Shang [20]	User-based collaborative filtering (CF) on Hadoop	Scalable because Hadoop is a cloud platform	Can't reduce recommendation response time for a single user,	Hadoop has been used, through which overcomes scalability problem (Big dataset can also be used)
Carlos E. Seminario , David C. Wilson [21]	CF using mahout, (User Based and Item Based)	Prediction accuracy for both user-based and item-based has been improved	Mahout similarity weighting is not very effective as a weighting techniques	Use of Mahout with Collaborative filtering has enhances the accuracy
Manos Papagelis, Dimitris Plexousakis [22]	User-based with implicit rating and explicit rating, Item-based with implicit rating and explicit rating	Prediction based on explicit rating is better than implicit rating	Scalability problem for big data	Item-based prediction algorithm is better than user-based algorithm
Trouong Khanh Quan, Ishikawa Fuyuki, Honiden Shinichi [23]	Clustering of items on stability of user similarity and apply CF	Good prediction accuracy	No. of groups must be given, Final clustering may be locally optimal	Clustering of item in a group performed better
Dhoha Almazro, Ghadeer Shahatah, et al. [24]	Cluster all items, Demographic information of user, Combine both user-based and item-based	Accuracy is 66.01%	Scalability for big data	Method has combined user-based and item-based
Hee Choon Lee, Seok Jun Lee, Young Jun Chung [25]	Neighborhood CF(NBCFA), Correspondence mean algorithm(CMA)			The preference prediction performance of CMA is better than NBCFA
Yanhong Guo, Xuefen Cheng, et al. [27]	CF based on trust factor, Cosine correlation and Pearson correlation has been used	CF based on trust factor is better than traditional CF	Scalability problem for big data	Trust factor is based on user who gives review for others
Kai Yu, Xiaowei Xu, Jianhua Tao, et al. [28]	Memory based CG, TURF1, TURF2, TURF3, TURF4	Reduces the storage requirement of training data	Scalability problem for big data	Select users with rational and novel profile
Dilek Tapucu, Seda Kasap, Fatih Tekbacak [29]	CF, Pearson correlation coefficient; Spearman corr. Coefficient; Tanimoto coefficient; Log likelihood similarity; Euclidean Distance Simm.		Scalability problem for big data, Improve quality is a challenge	Time complexity is less for Pearson Correlation coefficient
Mustansar Ali Ghazanfar, Adam Prugel Bnnett [30]	Method combines rating, feature and demographic information of item	Better prediction	Scalability problem for big data	Combining all factors provide better result
Badrul Sarwar, George Karypis, et al. [31]	Item-based CF	High quality recommendation	Scalability problem for big data	Item-based CF perform well for users who have rated less items.
Alexandros Karatzoglou, Alex Smola, Markus Weimer [32]	CF with hashing using-ϵ-intensive loss function, Huber loss function	Model can be scaled to bigger dataset on large server and to still dataset on small machines.	Time complexity	Hashing is used to bound the required memory, Loss function is used to achieve a large marginal model
Yajie Hu, Ziqi Wang, Wei Wu, Jianzhong Guo, Ming Zhang [33]	Semantic distance measurement and consider the features of movie. Recommendation based on YAGO and IMDB	Able to give a list of recommended movie along with stars of that movie.	Does not put user's feedback into consideration	Semantic distance has been used

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-10

Xiao Yan Shi, Hong Wu Ye, et al. [34]	Combine user based and item based	Better than traditional CF	Scalability problem for big data	To overcome scalability problem Cloud can be used.

References

[1] Andrew S Tanenbaum and Maarten van Steen, "Distributed Systems: Principle and Paradigms", in Pearson Prentice Hall, 2nd edition, may 2005.

[2] Jeffrey Dean and Sanjay Ghemawat, "Map-Reduce: Simplified data processing on large clusters", to appear in OSDI 2004.

[3] Chuck Lam, Hadoop in Action, Manning publication, 2010.

[4] S. Yu, D. Cai, J.-r. Wen, W.-y. Ma, Improving pseudo relevance feedback in web information retrieval using web page segmentation, in proceeding of the 12th international conference on world Wide Web www03 in new York, pp 11-13, ACM.

[5] Christian Kohlschttear and Wolfgang Nejdl, A Densitometric approach to web page segmentation, ACM 2008.

[6] David Fernandes, Edleno S. de Moura, Altigran S. de Silva, Berthier Ribeiro Neta and Edisson Braga, A site oriented method for segmenting web pages, SIGIR11, Beijing, ACM 2011.

[7] Jan Zeleny and Radek Burget, Cluster based page segmentation- A fast and precise method for web page perprocessing, ACM 2013.

[8] J. B. Schafers, J. Konstan and J. Riedi, Recommendation Systems in e-commerce, !st ACM conference on Electronic commerce ACM press, pp. 158-166, 1999.

[9] Sarwar B. and Karypis, Item based collaborative filtering algorithms in 10th International World Wide Web conference, pp 285-295, 2001.

[10] Mukund Deshpande and George Karypis, Item based top-N recommendation algorithm, in ACM Transactions Information Systems, volume 22, no. 1, pp 143-177, 2004.

[11] Aristomenis S. Lampropoulos and George A. Tsihrintzis, A survey approach to designing recommendation system , Springer 2013.

[12] Loren Terveen and Will Hill, Beyond recommender systems: Helping people help each other, In HCI In The New Millennium, Jack Aarrdl, Addison Wesley, 2001 page 2 0f 21.

[13] Christian Kohlschütter and Wolfgang Nejdl, A Densitometric approach to web page segmentation, ACM 2008.

[14] Erdinç Uzan, Hayri Volkan Agun, Tarik Terlikaya, A hybrid approach for extracting informative content from web pages, Science direct 2013.

[15] Fei Hu, Ming Li, Yi Nan Zhang, Tao Peng, Yang Lei, A non-template approach to purity web pages based on world density, Proceedings of International Conference on information engineering and application (IEA) 2012, Springer 2013.

[16] Zhao Cheng-Li and Yi Dong-Yun, A method for eliminating noises in web pages by style tree model and its applications, Wuhan university Journel of natural sciences 2004, vol-4, no. 5.

[17] Shekhar Babu Boddu, Eliminate The Noisy data from web pages using data mining techniques, GESJ: computer science and telecommunication 2013.

[18] Hiroyuki Sano, Shun Shiramatsu, Tadachika Ozono, A web page segmentation method based on page layouts and title blocks, IJCSNS international journal of computer science and network security, vol. 11 no. 10, 2011.

[19] Renato Dominguez Garcia, Alexandru Berlea, Philpp Scholl, Doreen Bhnstedt, Christoph Rensing, Ralf Steinmetz, Improving topic exploration in the blogosphere by detecting relevant segmentation, journal of universal computer science 2009.

[20] Zhi-Dan Zhao and Ming-Sheng Shang, User based collaborative filtering recommendation algorithm an hadoop, IEEE 2012.

[21] Carlos E. Seminario and David C. Wilson, Case study evaluation of mahout as a recommender plateform, presented in workshop on recommendation utility evaluation: Beyond RMSE, held in conjunction with ACM in Ireland, 2012.

[22] Manos Papagelis and Dimitris Plexousakis, Qualitative analysis of user based and item based prediction algorithms for recommendation agents, Science Direct 2005.

[23] Trouong Khanh Quan, Ishikawa Fuyuki, Honiden Shinichi, Improving accuracy of recommendation system by clustering item based on stability of user similarity, IEEE 2006.

[24] Dhoha Almazro, Ghadeer Shahatah, Lamia Albbulkarim, Mona Kherees, Romy Martinez, William Nzoukou, A survey paper on recommendation system, ACM 2010.

[25] Hee Choon Lee, Seok Jun Lee, Young Jun Chung, A study on improved collaborative filtering algorithm for recommendation system, IEEE 2007.

[26] Wu Yueping and Zheng Jianguo, A research of recommendation algorithm based on cloud model, IEEE 2010.

[27] Yanhong Guo, Xuefen Cheng, Dahai Dong, Chunyu Luo, Rishuang Wang, An improved collaborative filtering algorithm based on trust in e-commerce recommendation system, IEEE 2010.

[28] Kai Yu, Xiaowei Xu, Jianhua Tao, Martin Aster, Eans-Peter Kriegel, Instance selection techniques for memory based collaborative filtering, SIAM.

[29] Dilek Tapucu, Seda Kasap, Fatih Tekbacak, Performance comparision of cmbined collaborative filtering algorithm for recommender system, IEEE 2012.

[30] Mustansar Ali Ghazanfar and Adam Prugel Bnnett, A scalable, accurate hybrid recommender system.

[31] Badrul Sarwar, George Karypis, Joseph Konstan, John Riedl, Item based collaborative filtering recommendation algorithms, www10, in Hong Kong, ACM 2001.

[32] Alexandros Karatzoglou, Alex Smola, Markus Weimer, Collabrative filtering on a budget, Appear in proceedings of the 13th international conference on Artificial Intelligence and Statistics 2010, Italy.

[33] Yajie Hu, Ziqi Wang, Wei Wu, Jianzhong Guo, Ming Zhang, Recommendation for movies and stars using YOGA and IMDB, IEEE 2010.

[34] Xiao Yan Shi, Hong Wu Ye, Song Jie Gong, A personalized recommender integrating item based and user based collaborative filtering, IEEE 2008.

About Author (s):

Swati Pandey is pursuing M.Tech (CSE) in Amrita University, Coimbatore. She holds first rank in M.Tech. Her area of interest is Machine Learning and Distributed Computing.

Dr. T.Senthil kumar has around 12 years of teaching experience and 2 year of industry Experience. His area of interest includes cloud computing, software Engineering, Video processing, Wireless Sensor Networks, Dot Net Programming, JIST simulator, Data Mining. He is currently working as a Assistant Professor (Selection Grade) in computer science and Engineering Department at Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore. He has publication in 10 National Conferences and 6 International Conferences and 6 International journals.