ISSN: 0976-3104
SUPPLEMENT ISSUE

ARTICLE

A SURVEY PAPER ON COLLABORATIVE FILTERING BASED RECOMMENDER SYSTEMS IN BUSINESS INDUSTRIES

VE Jayanthi*, C Raja Kumar

[1]*Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul, INDIA*

[2] *Department of Information Technology, ULTRA College of Engineering & Technology for Women Madurai, INDIA*

ABSTRACT

Big Data hype gives the most attention to the recommender systems. Business industries are able to deliver value to their customers and to get significant revenue with the comfort of recommender systems. Recommender systems are in six main folds namely Collaborative Filtering (CF), Content-Based approach (CB), Demographic, Knowledge-based, Hybrid and Community-based approach. CF technique is a preferable approach to building successful recommender systems. So an extensive survey of the CF techniques is shortened for the researcher's betterment to accomplish their further work.

KEY WORDS
Recommender System, Collaborative Filtering, Item-based technique, User-based technique

INTRODUCTION

Inexpensive digital storages availability, mobile internet access facility, social media fad and IT devices innovation produces quintillion bytes of data every day. Such kind of data is very much worth for the business to predict customer needs and take right decision at the right time. A recommender system is a legitimate paradigm helps to the customer for their deserve choice. CF is one of the classifications under the recommender systems. The author coined the term collaborative filtering in an e-mail filtering system called Tapestry [1]. GroupLens automated recommendation system is proposed to provides personalized recommendations on Usenet postings [2].

Researchers continuously developed and implemented new version of CF based recommender systems. Even though the considerable amount of clarity on the existing CF methods and its works are not up to the mark. The performance of those methods is various with respect to the number of users, number of items, and sparsity level. Some of the methods perform decently in meager environments while others perform decently in complicated environments. Existing CF based recommender systems experimental result with respect to its parameters is analyzed. This paper is organized as introduction to various CF techniques algorithm and its challenges are in section II. The comparative study of CF methods is elaborated in Section III. Conclusion and future works are in section IV & V.

CF based technique and its challenges

Received: 15 Jan 2017
Revised: 14 Feb 2017
Published: 12 March 2017

CF techniques try to predict the additional items for a new user based on the list of favourite items previously rated by other users. There are two types of ratings. In the first type, ratings are explicitly asked from the user by a concrete rating scale. The second type gathers data implicitly from the user based on his/her accomplishment in a website. Captured through the user's actions and then the behaviour is analyzed to find the user's interest. This rating is easier than explicitly rating, but a user has more responsibility and it provides transparency in the rating process.

Tackle the highly sparse data, extreme scaling users and items counts are the challenges in high-quality predictions of data in CF technique's recommendations system. CF Techniques' should make reasonable recommendations in a short time period and to deal with additional problems like synonymy, shilling attacks, data noise, and privacy protection problems etc.,

CF algorithms are categorized into memory-based, model-based and hybrid collaborative filtering algorithms. Memory-based CF algorithms learn the rating matrix and find the recommendations based on the relationship between the request of user in a particular item and rating of the matrix. Based on user recommendations Model-based CF algorithm fits a parametric model to train the data to predict unobserved ratings and make recommendations. Hybrid CF algorithm is the combination of memory-based and model-based CF algorithms are investigated in the following section.

***Corresponding Author**
Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul, INDIA.

Memory-based CF algorithms

Memory-based CF algorithms exploit the entire user/item data loaded in the memory. Here the set of neighbors are found initially and then prediction is obtained. Distance measure is used to identify the neighbors. Similarity measure [0, 1], Dissimilarity measure [0, INF] are the two different approaches used to measure the distance between the users/items.

Similarity measures finds out the similarity between pairs of users or the pairs of items. Similarity measures are basically correlation based or vector cosine based measurement. Correlation based similarities measures are Pearson correlation and Spearman rank correlation. Pearson correlation is introduced and it is widely and successfully used as a similarity measure between users [2]. Constrained Pearson correlation similarly [3] is measure by considering user mean vote is the midpoint of the rating scale as a constant value to get high performance of Pearson correlation. Correlation based similarity measure is carried out to test the performance of various CF recommendation algorithms are described [2] and [4]. Cosine of the angle between two vectors are measured in the Cosine similarity measurement and it is described in [5] and [9]. Probability based similarity is the exceptional category of similarity measure. Conditional probability-based asymmetric similarity measure is explained [6]. Tanimoto coefficient brought in [7] is a similarity between two data sets and it is the ratio of intersections.

Dissimilarity measures are basically Euclidean distance which computes the distance between two objects in the Euclidean space. The Manhattan distance is based on Euclidean distance; calculate the distance by traversing the vertical and horizontal line in the grid base system. Metric of Euclidean space is Minkowski distance. Memory-based learning methods are highly sensitive to noise and redundancy. To remove the noise and redundancy in the memory based learning four techniques of Training User Reduction for Collaborative Filtering (TURF1-TURF4) is introduced [8].

Model-based CF algorithms

Machine learning and data mining techniques are used to find the patterns on the training data and make predictions for real time data to develop model-based CF algorithm. Model-based CF algorithm matches the model for the given rating matrix to issue the recommendations.

Bayesian models, clustering models, regression models, Markov decision processes (MDP) models and Latent Semantic models are proposed to fix the inadequacies of memory-based CF algorithms. Bayesian methods make a probabilistic kind model to solve the problem of inadequacies of memory in the CPU. Rule-based approach is used to find the correlation between the items in the association kind of model. Latent semantic analysis is used for MDP in aspect model based CF algorithms [10]. Multinomial mixture model in [5], user rating profile (URP) model in [11][35] are also use to overcome the fixing of inadequacies in memory-based CF algorithms. The classification algorithms are used in a model if the user ratings are categorical. Classifier is used for clustering method to groups the similar user in a class to evaluate the system. The regression and SVD methods are used if the user ratings are numerical in nature.

Probabilistic approach is enforced in model based algorithms and it is formulates the model from the user ratings on other items and predicts the expected value. Matrix factorization model is applied for extended stochastic gradient descent and alternating least squares (ALS) algorithms in [12] reduces the CF challenges. Bayesian Belief Net CF, Simple Bayesian CF, NB-ELR CF and TAN-ELR CF algorithms are developed based on Bayesian network model and they revelled [13] and [14].

Cluster model based CF are proposed with various algorithms, such as k-mean partitioning methods [15]. Density-based methods proposed [16] and extended in [17], data clustering methods [18]. Topic model based clustering algorithm [19] and [20]. The spreading activation model based CF algorithms [21] designed and implemented by the Leakey capacitor algorithm, Branch and bound serial symbolic search algorithm and Hopfield net parallel relaxation search algorithm.

Hybrid CF algorithms

The blended memory-based and the model based CF algorithms defeat the limitation of the native CF algorithms. Performance enrichment in the prediction of end result is achieved by hybrid based CF algorithm but the implementation is more expensive due to algorithm complexity.

Hybrid CF systems combine the CF techniques with another recommendation technique to make predictions or recommendations. Similarity measures play an important role of recommendation system. The above mentioned any two similarity measures combined to form a hybrid CF system. A few models are proposed to enrich a new hybrid CF system. Blended approaches are proposed to build a hybrid CFs to improve the prediction performance, overcomes the cold start problem and the sparsity problem.

Content-boosted CF algorithm is developed [22] by the combination of naive Bayes and a weighted Pearson correlation-based CF Algorithm. Combination of TANELR and Pearson correlation-based CF algorithms is used to developed a hybrid CF system is explained [23][33,34]. R. Burke, et al. 2002 [24] developed the Weighted hybrid recommender system, which is the combination of different recommendation techniques with their weights; accordingly it is named as mixed hybrid recommenders, cascade hybrid recommenders, and meta-level recommenders. Personality diagnosis algorithm in [25] groups the memory based and model based algorithms. Possible strategies to frame hybrid CF algorithms are considered and related in [26].

Cf system evaluation metrics

The eminence of a recommender system can be determined based on their assessment result. The type of metrics used depends on the type of CF techniques. Those metrics are broadly classified as predictive accuracy metrics, classification accuracy metrics and rank accuracy metrics. Mean Absolute Error (MAE) and its variations are the measurement of predictive accuracy metrics [27], [28], [30] and [31]. Classification accuracy metrics measures are precision, recall, F1-measure, and Receiver Operating Characteristic (ROC) sensitivity [21].

Pearson's product-moment correlation, Kendall's Tau, Mean Average Precision (MAP) are the parameters of rank accuracy metrics. hit-rate (HR) and the average reciprocal hit-rank (ARHR) are the measures in [29]. Most commonly-used CF metrics are MAE, Normalized Mean Average Error (NMAE), Root Mean Squared Error (RMSE), and ROC sensitivity.

Table 1: CF based recommendation systems analysis

CF categories	Techniques	Algorithm and Datasets	Performance Metrics	Implementation	Result and Remarks

Model based	Bayesian network, clustering, rule-based, MDP-based and Latent Semantic-based approaches	TyCo - Typicality based CF Algorithm
		
		Compared with:
		
		• Classical Base Line methods
		  ○ content-based (CB) - cosine similarity
		  ○ user-based CF - Pearson Correlation (UBCF)
		  ○ item-based CF - Pearson Correlation (IBCF)
		  ○ naive hybrid method
		  ○ CF with effective missing data prediction (EMDP)
		• state-of-the-art methods
		  ○ cluster based Pearson Correlation Coefficient method (SCBPCC)
		  ○ Weighted low-rank approximation (WLR)
		  ○ transfer learning-based collaborative filtering (CBT)
		  ○ SVD++
		
		Datasets: Movielens
		
		943 users, 1682movies, 1000,000 ratings
			MAE
			
			Coverage
				J2SE platform
				
				Pentium IV 3.2GHz
				
				2GB RAM
				
				Windows XP Professional
					Outperforms than
					
					many CF recommendation method.
					
					More accurate predictions with less number of big-error predictions.
					
					Issues addressed:
					
					Sparesity, Big-error in Prediction

		spreading activation algorithm compared with: • 3 Hop - Graph based CF • User based (correlation) • User based (vector similarity) • Item based Datasets: www.books.com.tw 9695 user nodes,2000 cust nodes,18771 links	Precision Recall f-measure rank score	Stored Procedure MySQL Python based sparse matrix library	Outperformed. Effectively alleviate the cold start problem. Over activation may dilute the data used to infer user performance **Issues addressed:** Sparesity
		Item-based top-N recommendation algorithm Compared with: User-based, Item-based Algorithms Datasets: real and synthetic datasets customer purchasing transactions, synthetic transaction dataset generator provided by the IBM Quest group	hit-rate, average reciprocal hit-rank	--	Accurate recommendations better than traditional user-based CF techniques. Designed to evaluate the effect of the similarity normalization. **Issues addressed:** Sparesity
Memory based	correlation based, Cosine Similarity based, Euclidean distance based	Enhanced Pearson Correlation Coefficient (PCC) algorithm Missing data prediction algorithm Compared with: state-of-the-art collaborative filtering approaches Datasets: Movielens 943 users, 1682movies, 1000,000 ratings	MAE	--	Outperforms than state-of-the-art collaborative filtering approaches. Combines users information and items information together. **Issues addressed:** Sparesity

Hybrid	content-based aspects into CF models (or) adding CF aspects into content-based models.	Combined User based, Item based approach Compared with : • Standard user based vector similarity • Item based adjusted cosine similarity • Cluster based Pearson Correlation coefficient • Aspect Model • Personality Diagnosis • User based Pearson correlation coefficient Datasets: Movielens 943 users, 1682movies, 1000,000 ratings	MAE	--	Outperformed. Generative probabilistic framework. **Issues addressed:** Sparesity

Comparision

Recommender systems are relatively enlightened compared with other research area in the field of information retrieval system. Research fascination in recommender systems has dramatically increased in recent days. Hence, would like to survey the existing CF based recommender systems techniques and its approaches in a broader way. Generally, CF based recommender systems outperform well compared to other legacy methods. Various techniques, corresponding dataset, technique related software, performance metrics and their merit and demerit of the existing CF algorithms are compared with state of art methods are listed in [Table 1].

It's content try to address the implementation issues of Sparesity and Big-error in Prediction issues present in the existing CF algorithm. The effectiveness of those approaches is evaluated experimentally using data from Movielens data set, the online book store data set, IBM Quest group's real and synthetic datasets. From [Table 1] model based CF algorithms gives the good performance as compared with the other types of CF algorithms is observed.

CONCLUSION

Memory-based, model-based CF and hybrid CF techniques of successful recommender systems are reviewed based on similarity metrics measured. Correlation between item based and user based are used to find the value of similarity metrics measures to evaluate the existing CF system. Memory-based CF algorithms are easy to implement and the performance decrease if the data are sparse in nature have good performances for dense datasets. Sparsity and scalability challenges are addressed in Model-based CF techniques. Hybrid CF techniques prediction performances is good but it is complicated. Recent days buying and selling through online dramatically increases huge volume of data, variety and its velocity. The model-based CF techniques are address the above said challenges in an easy way.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Goldberg D, Nichols D, Oki BM, Terry D. [1992] Using collaborative filtering to weave an information tapestry Commun ACM 35(12): 61–70

[2] P Resnick, N Iacovou, M Sushak, P Bergstrom, J Riedl. [1994] GroupLens: An Open Architecture for Collaborative Filtering of Netnews, In Proceedings of the Computer Supported Collaborative Work Conference.

[3] U Shardanand, P Maes. [1995] Social Information filtering Algorithms for Automating 'Word of Mouth', In Proceedings of CHI.

[4] Yi Cai, Ho-fung Leung, Qing Li, Huaqing Min, Jie Tang, Juanzi Li. [2013] Typicality-Based Collaborative Filtering Recommendation, IEEE Transactions on Knowledge and Data Engineering, 26 (3): 766 - 779

[5] J Breese, D Heckerman, C Kadie. [1998] Empirical analysis of predictive algorithms for collaborative filtering, in Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI '98).

[6] G Karypis. [2001] Evaluation of item-based top-N recommendation algorithms, in Proceedings of the International Conference on Information and Knowledge Management (CIKM '01), 247–254, Atlanta, Ga, USA.

[7] Toby S. [2007] Programming Collective Intelligence: Building Smart Web 20 Applications, O'reilly.

[8] Kai Yu, Xiaowei Xu, Jianhua Tao, Martin Ester and Hans-Peter Kriegel. [2002] Instance Selection Techniques for Memory-Based Collaborative Filtering, Proc Second SIAM International Conference on Data Mining (SDM'02)
[9] Sarwar, BG Karypis, J Konstan, J Riedl. [2001] Item-based Collaborative Filtering Recommendation Algorithms In Proc of the 10th International WWW Conference.
[10] T Hofmann and J Puzicha. [1999] Latent class models for collaborative filtering, in Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99), 688–693.
[11] B Marlin. [2003] Modeling user rating profiles for collaborative filtering, in Neural Information Processing Systems.
[12] Y Koren, R Bell, C Volinsky. [2009] Matrix factorization techniques for recommender systems, Computer, 42: 30–37.
[13] K Miyahara and MJ Pazzani. [2002] Improvement of collaborative filtering with the simple Bayesian classifier, Information Processing Society of Japan, 43(11).
[14] N Friedman, D Geiger, M Goldszmidt. [1997] Bayesian network classifiers, Machine Learning, 29(2-3): 131–163.
[15] MacQueen. [1967] Some methods for classification and analysis of multivariate observations, in Proceedings of the 5th Symposium on Math, Statistics, and Probability, 281–297, Berkeley, Calif, USA.
[16] M Ester, H-P Kriegel, J Sander, X Xu. [1996] A density-based algorithm for discovering clusters in large spatial databases with noise, in Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD).
[17] M Ankerst, MM Breunig, H-P Kriegel, J Sander. [1999] OPTICS: ordering points to identify the clustering structure, in Proceedings of ACM SIGMOD Conference, 49–60.
[18] T Zhang, R Ramakrishnan, M Livny. [1996] BIRCH: an efficient data clustering method for very large databases, in Proceedings of the ACM SIGMOD International Conference on Management of Data, 25: 103–114.
[19] T Hofmann. [1999] Probabilistic latent semantic analysis, in In Proc of Uncertainty in Artificial Intelligence, UAI, 289–296
[20] J Tang, J Zhang, L Yao, J Li, L Zhang, Z Su. [2008] Arnetminer: extraction and mining of academic social networks, in KDD '08 New York, NY, USA: ACM, 990–998
[21] Z Huang, H Chen, D Zeng. [2004] Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering, ACM Trans Inf Syst, 22(1): 116–142
[22] P Melville, R J Mooney, R Nagarajan. [2002] Contentboosted collaborative filtering for improved recommendations, in Proceedings of the 18th National Conference on Artificial Intelligence (AAAI '02), 187–192
[23] R Greinemr, X Su, B Shen, W Zhou. [2005] Structural extension to logistic regression: discriminative parameter learning of belief net classifiers, Machine Learning, 59(3) 297–322
[24] R Burke. [2002] Hybrid recommender systems: survey and experiments, User Modelling and User-Adapted Interaction, 12(4): 331–370
[25] DM Pennock, E Horvitz, S Lawrence, CL Giles. [2000] Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach, in Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI), 473–480
[26] Burke, R Hybrid. [2007] web recommender systems In: The Adaptive Web,377–408 Springer Berlin / Heidelberg.
[27] H Ma, I King, MR Lyu. [2007] Effective missing data prediction for collaborative filtering, in SIGIR: Proceedings of the 30th annual international ACM New York, USA: ACM, 39–46
[28] J Wang, AP de Vries, MJT Reinders. [2006] Unifying userbased and item-based collaborative filtering approaches by similarity fusion, in SIGIR ACM, 501–508
[29] M Deshpande, G Karypis. [2004] Item-based top-N recommendation algorithms, ACM Transactions on Information Systems, 22(1): 143–177
[30] K Yu, A Schwaighofer, V Tresp, X Xu, H-P Kriegel. [2004] Probabilistic memory-based collaborative filtering, IEEE Transactions on Knowledge and Data Engineering, 16(1): 56–69
[31] JL Herlocker, JA Konstan, LG Terveen, JT Riedl. [2004] Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems, 22(1): 5–53
[32] Kasnakoglu C. [2010] Control of nonlinear systerepresented by Galerkin models using adaptation-based linear parameter-varying models International Journal of Control, Automation and Systems,8(4): 748-761
[33] Stephygraph LR, Arunkumar N, Venkatraman V. [2015] Wireless mobile robot control through human machine interface using brain signals In Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), International Conference on, (IEEE), 596-603
[34] Arunkumar N, Ram Kumar K, Venkataraman V. [2016] Automatic Detection of Epileptic Seizures Using Permutation Entropy, Tsallis Entropy and Kolmogorov Complexity Journal of Medical Imaging and Health Informatics, 6(2): 526-531
[35] Arunkumar N, Kumar KR, Venkataraman V. [2016] Automatic Detection of Epileptic Seizures Using New Entropy Measures Journal of Medical Imaging and Health Informatics, 6(3): 724-730