Trends in Collaborative filtering Recommendation Technique

Nidhi Gupta
(Assistant Professor, GLNAIT, Mathura, India)
(nidhi0208@gmail.com)

Abstract: Recommendation system has proved as the key for successful online business. In this paper gives a study of the Collaborative filtering (CF) techniques used in the context of Recommender Systems. Study on various challenges faced by CF technique is discussed. Further, attempt is done to provide a comparative study of CF techniques, which surely helps various researchers in their work in the same area.
Keywords: Recommender systems, Collaborative filtering, cold start problem, memory based CF, model based CF.

I. INTRODUCTION

It is very common to say that 'Computers have changed the world'. With the rise in Internet revolution and advancement in the related field has contribute a lot in this area. Now, user centric approaches is used which predicts users interest on information, products and services among tremendous amount of available items by aggregating and analyzing suggestions from other users. Such a system is called a Recommender system which uses Data Mining and Machine learning approaches to predict the users interest.

There are two widely used approaches among recommender systems, Collaborative Filtering (CF) and content based filtering (CBF). The traditional task in the former one is to predict the utility of a particular item for the active user from the opinions of other similar users, and thereby make appropriate recommendations; on the other hand, later approach provides recommendations by comparing representations of content contained in an item to those of a user's interest ignoring opinions of other similar users [1, 2]. This paper provides a study of recommender systems, with the aim of imposing a degree of order on the diversity of the different aspects involved in their design and implementation. The first part provides a review of the state of the art of systems by providing the challenges faced and their possible solutions. The second part of the paper presents the comparison of different recommendation technique. The last part of the chapter discusses trends which might lead towards the next generation of systems.

II. CHALLENGES:

A. Lack of Data:

It refers to the situation when data is highly sparse. The data sparsity challenge appears in several situations, specifically, the cold start problem occurs when a new user or item has just entered the system, it is difficult to find similar ones because there is not enough information. The approach Nearest neighbor algorithms rely upon exact matches that cause the algorithms to sacrifice recommender system coverage and accuracy[1,13]. In particular, since the correlation coefficient is only defined between customers who have rated at least two products in common, many pairs of customers have no correlation at all [1]. Accordingly, Also, Pearson nearest neighbor algorithms may be unable to make many product recommendations for a particular user. This problem is known as reduced coverage, and is due to sparse ratings of neighbors.

Many solutions to deal with new user or cold start problem[2,3,4] are there. Recently a decision tree approach is used to deal with this problem in which users are rapidly interviewed by answering multiple choice questions. a linear regressor is learned within each node using all the previously obtained answers as input to predict item ratings.

Various approaches are there to deal with data sparsity problem. Dimensionality reduction techniques, such as Singular Value Decomposition (SVD)[5], remove unrepresentative or insignificant users or items to reduce the dimensionalities of the user-item matrix directly. The Latent Semantic Indexing (LSI) used in information retrieval is based on SVD [6], in which similarity between users is determined by the representation of the users in the reduced space. Goldberg et al. [7] developed eigentaste, which applies Principle Component Analysis (PCA), a closely-related factor analysis technique to reduce dimensionality. However, when certain users or items are discarded, some useful information for recommendations may get lost.

Hybrid CF algorithms, such as the content-boosted CF algorithm [8], are found helpful to address the sparsity problem, in which external content information can be used to produce predictions for new users or new items.

B. Scalability :

CF algorithms are able to search tens of thousands of potential neighbors in real-time, but the demands of modern E-commerce systems are to search tens of millions of potential neighbors as the demands or the users get increased the computational resources goes beyond the limits. Further, existing algorithms have performance problems with individual consumers for whom the site has large amounts of information. For instance, if a site is using browsing patterns as indications of product preference, it may have thousands of data points for its most valuable customers. These "long customer rows" slow down the number of neighbors that can be searched per second, further reducing scalability.

Dimensionality reduction techniques such as SVD can deal with the scalability problem and quickly produce good quality recommendations, but they have to undergo expensive matrix factorization steps. An incremental SVD CF algorithm [9] precomputes the SVD decomposition using existing users. Many techniques are there to deal with new ratings, without re- computing the low-dimensional model from scratch. Thus it makes the recommender system highly scalable.

Pearson correlation CF algorithm calculates similarities between all pairs of items, item-based Pearson CF calculates the similarity only between the pair of co-rated items by a user [10] hence able to achieve satisfactory results. A simple Bayesian CF[11] algorithm tackles the scalability problem by making predictions based on observed ratings . Model-based CF algorithms, such as clustering CF algorithms, address the scalability problem by seeking users for recommendation within smaller and highly similar clusters instead of the entire database [12], but there are tradeoffs between scalability and prediction performance.

C. Synonymy:

Synonymy refers to the tendency of same or similar items to have different names . Most recommender systems are unableto discover this latent association and thus treat these products diferently. For example, the seemingly different items "photo gallery" and "picture gallery " are actual the same item, but memory-based CF systems would find no match between them to compute similarity. Therefore, it decreases the recommendation performance of CF systems.

The SVD techniques, particularly the Latent Semantic Indexing (LSI) method, are capable of dealing with the synonymy problems. SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important ones .The LSI method gives only a partial solution to the polysemy problem, which refers to the fact that most words have more than one distinct meaning [19].

D. Gray Sheep.

Gray sheep refers to those users who think in a different way i.e their opinion about their liking and disliking are entirey different.and thus do not benefit from CF where as Black sheep are the opposite liking and disliking is nearly common to all group of people.

Claypool provided a hybrid approach combining content-based and CF recommendations by basing a predic-tion on a weighted average of the content-based prediction and the CF prediction. In that approach, the weights of the content-based and CF predictions are determined on a per- user basis, allowing the system to determine the optimal mix of content-based and CF recommendation for each user, helping to solve the gray sheep problem [20].

E. Shilling Attacks:

A shilling attack is an attack in which the system's recommendations for a particular item is manipulated by submitting misrepresented opinions to the system. The attack can have two objectives. Firstly, decrease the ratings of all the items outside its target item-set (push attack) to make them more recom-mended. Secondly, He may also increase the ratings (nuke attack) of other items to make its target item-set less recommended.

Two simple types of shilling attacks are RandomBot and AverageBot[14].

a) A RandomBot is a filterbot who randomly rate items outside of the target item-set with either the minimum rating (for nuke attack) or maximum rating (for push attack)[14].

b) An AverageBot is a filterbot where the rating is based on the average rating of each item following a normal distribution with a mean equal to the average rating for that item.

Another type of attack that may affect recommender are called Sybil attack in which a unauthorized user may create multiples users account in other to improve the recommendation of another user or another item. Recommender shall then provides ways to protect itself against those attacks since they are well known. Some systems provided CAPTCHA to stop filterbots from corrupting the ratings.[14].

ISBN: 9788175157538

Bell and Koren used a comprehensive approach to the shilling attacks problem by removing global effects in the data normalization stage of the neighbor-based CF, and working with residual of global effects to select neighbors. They achieved improved CF performance on the Netflix [17] data.

F. Evaluation of Recommender system:
Another major challenge is how to evaluate and compare between the recommender systems. Most of the sites uses the number of clicks as the evaluation criteria, but apart from this utility factor must be taken into account.

III. COMPARISON OF VARIOUS CF TECHNIQUES:

A. Memory based:
a) Recommendation or Prediction is based on similarity measure between user and item ratings.
b) Similarity measure includes Pearson correlation and vector cosine similarity. Also probabilistic method based on ranking similarity is also used.
c) It is highly dependent on human rating.
d) It is easy to implement but unreliable when data is sparse.

B. Model based:
a) Model refers to the data mining and Machine learning algorithm that uses user-item rating data to recommend or prediction e.g TAN-ELR(tree augmented naïve Bayes optimized by extended logistic regression)[15,16]
b) It uses Bayesian belief network, clustering, latent semantic technique as models. Also, Markov decision model is also used to produce much higher profit.
c) Prediction performance is high by compromising the scalability factor.
d) It uses dimensionality reduction technique for e.g. PCA, which further may lose some useful information.

C. Hybrid based:
a) It combines memory and model based
CF technique to overcome the predictions of other both technique e.g. Personality Diagnosis.
b) These algorithms such as content boosted algorithm are highly used to reduce the sparsity problem .

IV.CONCLUSION:

Recommender System have emerged as the most essential tool to deliver personalized recommendations for users in response to the above challenges particularly in E-commerce applications. Increasing number of users is affecting the Recommendation algorithms, therefore trend is changing with the new algorithms and new approach to deal with above mentioned challenges.

REFERENCES:

[1] Badrul M, Sarwar , George karypis, Joseph Konsta, John Ried "GroupLens ResearchGroup / Army HPC Research Center Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering" Currently with the Computer Science Department, San Jose State University, San Jose, CA 95112, USA

[2] G. Adomavicius and A. Tuzhilin, "Toward the next genera- tion 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.

[3] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, "Probabilistic memory- based collaborative filtering," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 56- 69, 2004.

[4] Mingxuan Sun, Fuxin Li, Joonseok Lee "Learning Multiple- Question Decision Trees for Cold-Start Recommendation", WSDM'13, February 4–8, 2013, Rome, Italy. ACM 978-1-4503-1869-3/13/02.

[5] D. Billsus and M. Pazzani, "Learning collaborative infor-mation filters," in Proceedings of the 15th International Conference on Machine Learning (ICML '98), 1998.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis,"Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[7] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigen-taste: a constant time collaborative filtering algorithm," Information Retrieval, vol. 4, no. 2, pp. 133-151, 2001.

[8] P. Melville, R. J. Mooney, and R.Nagarajan, "Content- boosted Collaborative filtering for improved recommenda- tions," in Proceeding of the 18th National Conference on Artificial Intelligence(AAAI '02), pp.187-192, Edmonton, Canada, 2002.

[9] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremen-tal SVD-based algorithms for highly scaleable recommender systems," in Proceedings of the 5th International Conference on Computer and Information Technology (ICCIT '02), 2002.

[10] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item- based collaborative filtering recommendation algorithms," in Proceedings of the 10th International Conference on World Wide Web (WWW '01), pp. 285-295, May 2001.

[11] K. Miyahara and M. J. Pazzani "Improvement of collabora- tive filterin with the simple Bayesian classifier,"Information Processing Society of Japan, vol. 43, no. 11, 2002.

[12] G.-. Xue, C. Lin, Q. Yang, et al. "Scalable collaborative filtering using cluster-based

ISBN: 9788175157538

smoothing," in Proceedings of the ACM SIGIR Conference, pp. 114-121, Salvador, Brazil, 2005.

[13] Konstan, J., Miller, B., Maltz, D., Herlocker, J.Gordon, L., and Riedl, J. (1997). GroupLens:Applying Collaborative Filtering to Usenet News.Communications of the ACM, 40(3), pp. 77-87.

[14] Dhoha Almazro, Ghadeer Shahatah, Lamia Albdulkarim "A survey paper on Recommender system" arXiv:1006.5278v4 [cs.IR] 24 Dec 2010.

[15] X. Su and T. M. Khoshgoftaar, "Collaborative filtering for multi-class data using belief nets algorithms," in Proceedings of the International Conference on Tools with Artificial Intelli- gence (ICTAI '06), pp. 497-504, 2006.

[16] R. Greinemr, X. Su, B. Shen, and W. Zhou, "Structural extension to logistic regression: discriminative parameter learning of belief net classifiers," Machine Learning, vol. 59, no. 3, pp. 297-322, 2005.

[17] Netflix prize, http:// www.netflixprize. com/.

[18] Y. Koren, "Tutorial on recent progress in collaborative filtering," in Proceedings of the the 2nd ACM Conference on Recommender Systems, 2008.

[19] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[20] M. Claypool, A. Gokhale, T. Miranda, et al., "Combining content-based and collaborative filters in an online newspa- per," in Proceedings of the SIGIR Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, Calif, USA, 1999.

ISBN: 9788175157538