A Survey of Recommender System from Data Sources Perspective

Huaiyu Pi[1, a], Zhenyan Ji[1,b*] and Chun Yang[1,c]
[1] School of Software Engineering, Beijing Jiaotong University, Beijing, China
[a]16121731@bjtu.edu.cn, [b]zhyji@bjtu.edu.cn, [c]17121715@bjtu.edu.cn

Keywords: Recommender system; Rating; Text; Social network; Multiple data

Abstract. In order to solve the problem of information overload in big data era, the personalized recommender system has been widely used. Collaborative filtering, as a classical algorithm, has become the basis of the recommender system. In recent years, there are more and more recommender systems based on multiple data sources are proposed. Today's recommender systems integrate multiple data sources and recommendation methods are more accurate and explainable compare with rating-based recommendation systems. How to integrate multiple data sources to further improve the accuracy and interpretability of recommendation results, reduce computational complexity and cold start risk has become the key content of recommendation researches.

Introduction

With the advent of the era of big data, the recommender system is widely used in e-commerce, news, video and music areas to help users find valuable information quickly in massive data. In the mid-1990s, collaborative filtering is applied to filtering mail and news. The recommendation system can not only recommend the list of interest to the user, but also help users make decisions to maximize user profits or increase the profit of the businesses. Up to now, major international online shopping or media platforms such as Taobao, Amazon and YouTube are all integrate recommendation algorithms.

Due to the differences in user hobbies, areas of concern, personal experiences, etc., different user have different needs. While the non-personalized recommendation system recommends the same items for all users, and cannot accurately recommend the items that users like, the personalized recommendation system in which different users can get different recommendation content as an important feature has received great attention. This paper focus on the personalized recommender systems and classifies them according to the data source or type they use. In the early stage, the recommender systems mainly use single data such as ratings to predict user preferences. The rating-based methods are not accurate because ratings contain limited information. In order to reveal the preferences of users, multiple data and integrated models are applied to recommender systems.

This article is divided into five sections, section 2 is an introduction of basic recommendation methods, section 3 focus on the problems in the recommender systems, section 4 is discussing the recommender systems using multiple data, section 5 is some problems and advanced topics.

Basic Recommendation Methods

In the early stage, the recommender systems use single method to predict user preferences. These recommender systems always predict the ratings that users may give to the items. Some of them only give the top k items for the users based on the similarity regardless of the ratings. The methods of recommender system are mainly divided into four categories[1,2,3]: demographic-based recommendation, content-based recommendation, collaborative filtering-based recommendation and hybrid recommendation. Up to now, most of the recommender systems are use multiple recommendation methods to get better results. But it is still significant for us to analysis the feature of every single recommendation methods.

The demographic-based method uses the user's primarily information, including gender, age, and so on. This method make recommendation based on the similarity of the users. But people with the

same gender and age may have different preferences. The content-based method is similar to the demographic-based method, because it also compares the similarity. The difference is content-based method make use of the item information to select the most similar items. CF(Collaborative filtering) is the most widely used method in recommendation systems. Different from the previous methods, It uses rating history to calculate the user similarity and the item similarity. Therefore, recommender system make recommendation based on the similarity of users or items. The hybrid recommendation method combines multiple existing methods to gain better result.

Recommendation Problems

Some challenges have emerged during the process of recommender system development. When a new user comes to the system, he has no record or few records. It is different for recommender algorithms to predict user preferences. This problem is called cold start. The similar problem has been found in items when a new item is added to the system. The other two problems are explanatory and diversity. If the recommended items are rich in variety and given reasonable explanations, the users will no doubt be more convinced of the recommendation results. In addition, the user privacy and information security are also important. In order to solve the problems, we can improve recommendation results with the help of multiple types of additional information. Obviously, we also need to design more algorithms to make use of multiple data.

From Single to Multiple Data Recommendation

Rating Recommendation. The earliest recommender systems mainly make recommendations based on users' explicit or implicit ratings. Konstan et al.[4] considered the variety and big volume of news, he applied the CF to news recommender system that provides personalized news based on users' ratings in 1992. This system called Usernet news provide users an effective news recommendation service. With the increase of users, the user-based CF faces with the challenge of high computational complexity, but the item number is growing slower than the user number in many web applications. Deshpande et al.[5] proposed item-based CF algorithms to decrease the computational complexity up to two orders of magnitude. It calculates item similarity and makes the recommendation based on item similarity. It can also improve the recommendation quality based on implicit feedbacks. When the number of users and items is very large but the number of rating is small, the computational cost of user-based and item-based recommendation method become high and the result is inaccurate. Salakhutdinov et al.[6] proposed a Probabilistic Matrix Factorization (PMF) model which is not only recommend accurate result but scales linearly with the increase of user and item numbers. The final experiment result shows the PMF model with Restricted Boltzmann Machines recommend better results than the Netflix recommender system.

Rating and Text Recommendation. Ratings are simple and contain limited information, while reviews contain richer information. To increase the recommendation accuracy, many researchers have done researches about recommendation based on ratings, review texts and contextual text and so on [7]. Some researches analysis the ontology or build library to obtain the opinion of users and the feature of products. Aciar et al.[8] make recommendation based on the ontology and its weights for every reviewer. Takuma et al. [9] implement an analysis engine called "MeCab" to extract feature words from user reviews to extend the base word library. The similarity between users is calculated in account of feature words to find the most relative users. The more similar between the two users, the more weights are given. The final score to the hotels will be predicted based on the rating given by the reviewers and the weights which represent the similarity. In order to mining common features from different texts, topic model is used. Xu et al.[10] extract the topic words from proposals and find the experts profile by text mining. Then the topics and profiles are classified to represent the different feature of experts and proposals. Finally, several experts are recommended to every proposal based on the similarity. The previous methods focus on the similarity comparison, but McAuley et al.[11] combined latent factor from ratings and text reviews to improve the accuracy and Interpretability. The

proposed HFT model combine topic model and matrix factorization model to provide recommendation and topics. The model test in 42 million reviews and 10 million users significantly improved accuracy. Zhang et al. [12] proposed Explicit Factor Model(EFM) which regard the feature extract from review text as explicit factor. The paper considers both of the explicit features extract on phrase-level and the latent factors factorize from the rating matrix as influence of the recommendation results. An online website is implemented to infer that the explanations generated by the feature words have great influence on user decisions. Besides topic model and phrase analysis, word embedding and neural network encoder is leveraged to get the representation of texts. Bansal et al. [13] presented an end-to-end neural network model to make scientific paper recommendation. The deep recurrent neural network is applied to encoder the text and the gated recurrent units are used to undertake the CF tasks. The relationship between users or items is also hidden in the text. Xu et al. [14] use text mining technology to find the hidden community of users and the hidden group of items. Because different people concern about different topics and different items have different attributes, the recommendations will be more effective depending on the user and item clusters.

Social Network and Other Recommendation. A survey in early 2000s shows people believe in the recommendations from friends more than online recommender systems [15], friendship has significant influence in user decision. Many researchers focus on the mechanism that how trust relationship affect the recommendation results. Models for calculating trust values in a social network can be categorized into three main types: link prediction, community detection and matrix factorization.

The link prediction method predict the possibility of different people connecting in the future[16]. Jamali and Ester[17] proposed TrustWalker to calculate the trust value between two unconnected users to increase the accuracy of the trust-based method. Furthermore, the TrastWalker model can provide the confidence of recommendation results. To processing the binary and continuous value networks, Golbeck et al [18] proposed two sets of algorithms to predict connection in binary and continuous social networks respectively. The proposed model gains better result compare to CF based methods when test in the film website and mail application .

There are already many community discovery algorithms, but how to combine the community and recommendation method is a problem. Zhang et al. [19] presented a semi-supervised model combine topic and social networks called TopRec. This model use topic to filter out topical community to recommend items that user interested in. The experiment in real world datasets approves the effetiveness of this model.

The matrix factorization method achieves high accuracy in many state-of-art recommender systems. Therefore, many models employ it to help the analysis of social networks. Purushotham et al. [20] leverage topic modeling and social matrix factorization to combine text and social networks. He et al. [21] try to utilize friendship to find the behavior of user in decision making. The experiments based on the analysis on social network and product rating shows users have similar decisions and preferences. The topic filtering is also applied in this model to obtain more accurate social relationships and improve the recommendation accuracy. The trust prediction and matrix factorization are integrated in some researches. Jamali et al. [22] integrate the trust propagation and matrix factorization to build a model-based approach for recommendation. The experiment shows the integrated model improves the accuracy especially in cold start cases. To deal with the data sparsity and scalability problem in recommender systems. Ma et al. [23] proposed SoRec model which make use of PMF to analyze the ratings and the social networks. The accuracy of SoRec model is very high even compares to some state-of-art approaches. Both of the explicit and implicit feedback indicate the user's preferences. Guo et al. [24] extend the SVD++ model with trust information to build a trust-based matrix factorization model. The proposed TrustSVD model improve the accuracy of rating prediction tasks compare to ten recommendation models.

Conclusions

With the increasing of applications in the Internet, the source of data is getting more and more richer. Therefore, the various factors in the new data brings new challenges. It is also a chance to create novel methods to achieve better recommendation results. Social networks are still the focus of the recommendation research, integration methods and new algorithms will continue to appear in the future. The sound, location and other user preference information are received more and more attention. We believe that the future of the recommender system will be a hot area of innovation and research.

Acknowledgements

Supported by the Fundamental Research Funds for the Central Universities (2017YJS215).

References

[1] Pazzani M J. A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review, 1999, 13(5-6):393-408.
[2] Gunawardana A, Shani G. *A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*. (JMLR.org, 2009).
[3] Melville P, Mooney R J and Nagarajan R. *Content-boosted collaborative filtering for improved recommendations* (Eighteenth national conference on Artificial intelligence. American Association for Artificial Intelligence, 2002), p.187
[4] Konstan J A, Miller B N, Maltz D, et al. GroupLens: applying collaborative filtering to Usenet news. Cacm, 1997, 40(3):77-87.
[5] Deshpande M, Karypis G. *Item-based top- N, recommendation algorithms* (ACM, 2004).
[6] Salakhutdinov R, Mnih A. *Probabilistic Matrix Factorization* (International Conference on Neural Information Processing Systems. Curran Associates Inc., 2007), p.1257
[7] Domingues M A, Sundermann C V, Manzato M G, et al. *Exploiting Text Mining Techniques for Contextual Recommendations* (Ieee/wic/acm International Joint Conferences on Web Intelligence. IEEE, 2014), p.210.
[8] Aciar S, Zhang D, Simoff S, et al. Informed Recommender: Basing Recommendations on Consumer Product Reviews. IEEE Intelligent Systems, 2007, 22(3):39-47.
[9] Takuma K, Yamamoto J, Kamei S, et al. *A Hotel Recommendation System Based on Reviews: What Do You Attach Importance To?* (International Symposium on Computing & NETWORKING. IEEE, 2016), p.710.
[10] Xu Y, Zuo X. *A LDA model based text-mining method to recommend reviewer for proposal of research project selection* (International Conference on Service Systems and Service Management. IEEE, 2016), p:1.
[11] Julian J. McAuley, Jure Leskovec. *Hidden factors and hidden topics: understanding rating dimensions with review text* (Proceedings of the 7th ACM Conference on Recommender System. New York: ACM, 2013), p.165-172.
[12] Yongfeng Zhang, Guokun Lai, Min Zhang, et al. *Explicit factor models for explainable recommendation based on phrase-level sentiment analysis* (Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2014), p.83.
[13] Bansal T, Belanger D and Mccallum A. Ask the GRU: Multi-Task Learning for Deep Text Recommendations. 2016:107-114.
[14] Yinqing Xu, Wai Lam and Tianyi Lin. *Collaborative Filtering Incorporating Review Text and Co-clusters of Hidden User Communities and Item Groups* (Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM), p.251.

[15] Sinha R R, Swearingen K. *Comparing Recommendations Made by Online Systems and Friends* (DELOS workshop: personalization and recommender systems in digital libraries, 2001), p.106.
[16] Samanta S, Pal M. Link prediction in social networks. Springerbriefs in Computer Science, 2016:246-250.
[17] Jamali M, Ester M. *TrustWalker : a random walk model for combining trust-based and item-based recommendation* (ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009), p.397.
[18] Golbeck J A. *Computing and applying trust in web-based social networks*, 2005.
[19] Zhang X, Cheng J, Yuan T, et al. *TopRec:domain-specific recommendation through community topic mining in social network* (International Conference on World Wide Web. ACM, 2013), p.1501.
[20] Purushotham S, Liu Y and Kuo C C J. *Collaborative topic regression with social matrix factorization for recommendation systems* (International Coference on International Conference on Machine Learning. Omnipress, 2012), p.691.
[21] He J, Chu W W. A Social Network-Based Recommender System (SNRS). 2010, 22(3):47-74.
[22] Jamali M, Ester M. *A matrix factorization technique with trust propagation for recommendation in social networks* (Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010), p.135.
[23] Ma H, Yang H, Lyu M R, et al. *Sorec: social recommendation using probabilistic matrix factorization* (Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008), p.931.
[24] Guo G, Zhang J, Yorke-Smith N. Trustsvd: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings. 2015