Advanced Science and Technology Letters
Vol.51 (CES-CUBE 2014), pp.32-35
http://dx.doi.org/10.14257/astl.2014.51.08

Similar User Clustering based on MovieLens Data Set

Haesung Lee[1] and Joonhee Kwon[1]

[1] Department of Computer Science, Kyonggi University,
San 94-6, Yiui-dong, Yeongtong-ku, Suwon-si, Gyeonggi-do, Korea
{seastar0202, kwonjh}@kgu.ac.kr

Abstract. General recommender algorithms compose persomalized recommendrations based on similar users. In this paper, we present new social clustering method. Based on this method we cluster similar users belonging to the social recommender network. The social recommender network is generated from the data set of MovieLens. Through the presented similar user clustering method, we effectively compose similar user clusters. Consequentially, the proposed method help recommender systems efficiently compose personalized recommendations.

Keywords: Recommender system, similar users, clustering, social network.

1 Introduction

Personalized recommendations are especially important in e-commerce sites where the variety of choices is large, the taste of the customer is important. Some of the major e-commerce sites, like Amazon and Netflix, successfully apply recommender systems to deliver automatically generated personalized recommendation to their customer. One of the earliest and most successful recommender technologies is collaborative filtering (CF) [1]. CF has been very successful in both research and practice.

Many recommender techniques have been developed in the past decade, but a considerable amount of them were constructed with small datasets and they are entirely unrealistic attempts. While the tremendous growth of these data sets in recent years poses some key challenges, several recommender systems suffer from performance and scalability problems when dealing with larger datasets.

In this paper, we present new social clustering method. Based on this method we cluster similar users belonging to the social recommender network. The social recommender network is generated from the data set of MovieLens. Through the presented similar user clustering method, we effectively compose similar user clusters. Consequentially, the proposed method help recommender systems efficiently compose personalized recommendations.

ISSN: 2287-1233 ASTL

2 Related work

Social network theory can be used to model such a recommendation system of people versus items as an affiliation network and distinguishes between a primary mode and a secondary mode, where a mode refers to a distinct set of entities that have similar attributes [2]. In the view of recommender systems, the primary mode could be regarded as users. And, the secondary mode is considered as items. In other words, in a recommender system, the rating patterns of people on items induce an implicit social network and influence connectivities in the network.

Clustering is a kind of data mining technique for discovering interesting patterns from a given database [3]. The main idea of clustering is that given n data pointing in a m-dimensional metric space is divided into k clusters so that all the data pointing within one cluster has a closer similarity than the data within any other cluster [4]. Once the clustering is complete, the performance of recommender systems can be very good, since the size of data that must be analyzed much smaller. Although, the clustering result may depend on the initial seeds, however, there are few studies for the mechanism which optimize the initial seeds. Ultimately, clustering techniques usually produce less-personal recommendation than other methods and most often lead to worse accuracy than CF algorithms.

3 Similar user clustering based on the data set of MovieLens

Based on the argument of [5], we generate a social network with the recommender dataset. With the generated social network, we compose the similar user index. The similar user index is used for improving the performance of recommender systems.

The MovieLens dataset is composed of user's profile, item's metadata and the rating for items. The total size of data set is about 1Mega byte.

We understand the fact that there are relationships between a user and an item on which the user assign the rating value with their preference. According to the social network theory [6], we consider the user as the primary mode and the item as the secondary mode. Based on this concepts, we deduct social network model from the recommender dataset. A recommender dataset $\mathcal{R}$ can be represented as a bipartite graph $G = (U \cup I, E)$ like shown Fig. 1(a), where U is the set of people, I is the set of items, and the edges in E represent the ratings of items.

For improving the accuracy of clustering which assigns similar user in the same cluster, we take three phases in the clustering. Therefore, there are many studies about the selection of initial seeds to increase the accuracy of clustering. In the domain of recommender systems, however, there are few studies which consider the feature of the recommender systems. In this paper, we propose new seeds selection mechanism for improving the accuracy of clustering similar users. In order to choice seeds of each cluster, we use the social network generated from the dataset of MovieLens. We assume that influential users node in the social networks have much more relationships with other users. In other words, the influential user node could be considered as a seed of each similar user cluster because the influential user node have higher possibility to take relationships with any other user nodes.

Similar user clustering techniques work by identifying groups of users who appear to have similar preferences. Each seed which is selected in the phase 1 is located in the center of the cluster. For measuring the similarity, we use the square of the Euclidean distance measurement.

In order to improve the quality and the accuracy of the similar user clustering, we take clustering refinement technique at the final step of the similar user clustering. For this, we apply the modularity technique. In the domain of information clustering, the modularity is a criterion for evaluating the quality of partitioning a network into clusters [7]. While there are various modularity technique, Q is widely known as the most accurate [8].

4 Conclusion

In this paper, we present new social clustering method. Based on this method we cluster similar users belonging to the social recommender network. The social recommender network is generated from the data set of MovieLens. For the social clustering, we consider the concept of the affiliation network and construct a social network from the recommender data set. Based on the generated social network, the similar user clustering is performed. Through the presented similar user clustering method, we effectively compose similar user clusters. Consequentially, the proposed method help recommender systems efficiently compose personalized recommendations.

Acknowledgments. This work was supported by the Gyonggi Regional Research Center (GRRC) and Contents Convergence Software (CCS) research center.

References

1. Bobadilla, J., Ortega, F.: A. Hernando, and A. Gutierrez. Recommender systems survey. Konwledge-Based Systems, 46(0), pp. 109-132 (2013)
2. Schwartz, M.F. and Wood, D.C.M.: Discovering Shared Interests Using Graph Analysis, Communications of the ACM, Vol. 36(8), pp.78-89. (1993)
3. Dubes, R.C., Jain, A.K.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs. (1988)
4. Hartigan, J. A., & Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm. Applied statistics, 100-108. (1979)
5. Mirza, B. J., Keller, B. J., & Ramakrishnan, N.: Studying recommendation algorithms by graph analysis. Journal of Intelligent Information Systems, 20(2), 131-160. (2003)
6. Robalino, D. and Gibney, M.: A model of the Impact on Movie Demand of Social Networks and Word of Mouth Recommendation, URL:http://www.metaculture.net/Searches/movies. (1999)

7. Newman, M. EJ, and Girvan, M.: Finding and evaluating community structure in networks. Physical review E 69, no.2 (2004)
8. Xu, X., Yuruk, N., Feng, Z., and Schweiger, T.: SCAN: a structural clustering algorithm for networks. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 824-833. ACM. (2007)

