Hybrid Recommender System using Semi-Supervised Clustering based on Gaussian Mixture Model

Yihao Zhang
College of Computer Science and Engineering,
Chongqing University of Technology,
Chongqing, China
E-mail: yhzhang@cqut.edu.cn

Xiaoyang Liu
College of Computer Science and Engineering,
Chongqing University of Technology,
Chongqing, China
E-mail: lxy3103@cqut.edu.cn

Wanping Liu
College of Computer Science and Engineering,
Chongqing University of Technology,
Chongqing, China
E-mail: wpliu@cqut.edu.cn

Changpeng Zhu
College of Computer Science and Engineering,
Chongqing University of Technology,
Chongqing, China
E-mail: is99zcp@cqut.edu.cn

***Abstract*—Recommender systems are used to make recommendations about products, information, or services for users. Most existing recommender systems implicitly assume one particular type of user behavior. However, other recommender system utilizes different particular type information by combining different techniques to improve the quality of the recommendation. This paper proposed a novel personalized recommendation method that utilizes semi-supervised clustering based Gaussian mixture model, which provides a hybrid recommender method by combining demographic method and user-based collaborative filtering method. The result from various simulations using MovieLens data set shows that the proposed recommender method performs better and helps to improve the quality of recommendation rating.**

Keywords—hybrid recommender systems; semi-supervised clustering; Gaussian mixture model

I. INTRODUCTION

The amount of data in our world has been exploding, which makes it difficult for the decision-maker to identify useful information. In order to get the most value out of their data, the challenge is to ensure the right information is getting to the right employees, because the large number of information make it difficult for users to be aware of them or even look through them. The personalized recommender systems are important applications that can address this problem and suggest items that suit the user's needs [1]. Typically, a recommender system compares the user's profile to some reference characteristics, and seeks to predict the rating that a user would give to an item they had not yet considered [2]. The key element of a recommender system is the user model that contains knowledge about the individual preferences, which determine his or her behavior in a complex environment.

Recommender systems are known to address the information overload problem and reduce search effort [3], and they provide people with suggestions for items which are likely to be of interest for them. Usually, there are three main types of recommendation methods: collaborative filtering [4], content-based [5] and hybrid method [6]. Recent work on hybrid recommender systems has shown that recommendation accuracy can be improved by combining multiple data modalities and modeling techniques within a single model [7][8][9]. Hybrid recommender method Commonly uses a combination of collaborative filtering with demographic filtering [10] or collaborative filtering with content-based filtering [11][12] to exploit merits of each one of these techniques. Hybrid recommender method is usually based on biologically inspired or probabilistic such as genetic algorithms [13][14], neural networks ,Bayesian networks[16] and clustering[17]. In this paper, we propose a hybrid recommender method using semi-supervised clustering based on Gaussian mixture model (SSCGM), which leverages the information of user behavior information and user demographic information in order to build adaptable and extensible hybrid recommender systems. In particular, distance metric is used to measure the similarity of user preferences and Gaussian distribution information is used to measure the similarity of user behavior.

The rest of this paper is structured as follows. The section 2 introduces the different various types of input data and the workflow of hybrid recommender system. The section 3 describes semi-supervised clustering based Gaussian Mixture Model. The section 4 illustrates experimental result and analysis. Finally, we provide some concluding remarks and suggestions for future work in Section 5.

II. HYBRID RECOMMENDER SYSTEM WITH SEMI-SUPERVISED CLUSTERING

The heart of recommendation technologies is the algorithms for making recommendations based on various types of input data. The possible input data of recommender system is varied, but they can be grouped into three levels (user, items and review), which correspond to rows, columns, values in the matrix. Item Profile is used to describe the properties of an item, and usually it is different according to different item. For the book recommendation, item profile likely to include books category, author, number of pages, the time of publication, publishers; for news recommendation, item profile is possible that the news text, keyword ,time and so on;

978-1-5090-2303-5/16 $31.00 © 2016 IEEE
DOI 10.1109/CW.2016.32

for the film, it may be the title, duration, release time , actor, plot description. User Profile is used to describe a user's personality, which may have different representation and similar to item profile, such as the user's gender, age, income, active time, city and so on. Review is a contact with user and items, such as use an integer of 1 to 5 to indicate the extent of the user preferences to items, and may also contain a lot of different information, such as the user reviews text to items, the view history of user, the purchase history of user, etc.

A hybrid recommender system is proposed in this paper in order to fusion the demographic characteristic and user behavior information for improve recommendation quality. The input includes the user id (u_{id}), the total recommended number (N). The pseudo code description of the hybrid recommender algorithm is shown as Fig.1, and it essentially has five steps.

Step One. In the initialization stage, if there is not any history data of the browse items, a random algorithm is used to deal with the cold-start problem of recommender system.

Step Two. We calculate the distance of u_{id} and the other u_i according to user's demographic information, and use a function to convert $dis(u_{id},u_i)$ into weight matrix as a regularizer used in constructing the Gaussian mixture model.

Step Three. We build the Likelihood of Gaussian distribution to estimate users' similarity by utilizing user's behavior information, and then combine the regularizer into Gaussian mixture model for constructing the objective function of semi-supervised clustering algorithm, and then to solve it to build the clustering model.

Step Four. We calculate the similarity of users by using semi-supervised clustering algorithm, and get k most similar user using a mechanism.

Step Five. We recommend the top N items of browsed items by similar user according to some recommendation rule, and output the recommendation result.

Input: u_{id} , N

OutPut: recommender items (T_r)

Algorithm: *Hybrid recommender algorithm*

1. Initialization

T_b = the browsed items by u_{id} ;

T_{all} = the browsed items of all users;

if (T_b =0 and T_{all} = 0) then

T_r = recommended N items by random algorithms;

return T_r ;

end if

2. Distance metric learning using demographic characteristics

$U = \{u_1, u_2, \cdots, u_m\}$;

for each (u_i in U) do

$dis(u_{id},u_i)$ = calculate the distance of u_{id} and the other u_i according to user's demographic information ;

end for

$w_{ij} \leftarrow f(dis(u_{id},u_i))$, convert $dis(u_{id},u_i)$ into weight matrix by function;

3. Likelihood of GMM using user's behavior information

Combine regularizer with Gaussian mixture model

$$\ell = \sum_{i=1}^{m}\sum_{l=1}^{k} P(c_l \mid x_i)(\log p(x_i \mid c_l;\mu,\Sigma)+\log\Phi_l) - \lambda R ,$$

where $R = \sum_{i,j=1}^{m} D_{ij}W_{ij}$;

4. **Similarity calculation based on semi-supervised clustering**

sim_u_{id} = the number of k most similar user ;

5. **Recommended N items by clustering algorithm**

T_r = recommended top N items of browsed items by sim_u_{id} ;

Return T_r ;

Fig.1. Pseudo code description of the hybrid recommender algorithm

III. Semi-Supervised Clustering based on Gaussian Mixture Model

Suppose $P_i(c)$ and $P_j(c)$ denote the two distributions in Gaussian mixture model, the KL-Divergence between them can be defined as formula (1):

$$D(P_i(c) \,\|\, P_j(c)) = \sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)} \tag{1}$$

For the symmetric equation, we usually use formula (2) to measure the similarity between distributions $P_i(c)$ and $P_j(c)$.

$$\begin{aligned} D_{ij} &= \frac{1}{2}(D(P_i(c) \,\|\, P_j(c)) + D(P_j(c) \,\|\, P_i(c))) \\ &= \frac{1}{2}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)} + \sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)}) \end{aligned} \tag{2}$$

Let $P_i(c) = P(c \mid x_i)$, the smoothness of the conditional probability $P(c \mid x)$ can be defined as formula (3):

$$\begin{aligned} R &= \sum_{i,j=1}^{m} D_{ij}W_{ij} \\ &= \frac{1}{2}\sum_{i,j=1}^{m}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)} + \sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})W_{ij} \end{aligned} \tag{3}$$

While incorporating the above smoothness term into the likelihood of original GMM, we can get new objective function be defined as formula (4):

$$\begin{aligned} \ell_{new} &= \ell - \lambda R \\ &= \sum_{i=1}^{m}\sum_{l=1}^{k} P(c_l \mid x_i)(\log p(x_i \mid c_l;\mu,\Sigma)+\log\Phi_l) \\ &\quad - \frac{\lambda}{2}\sum_{i,j=1}^{m}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)} + \sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})W_{ij} \end{aligned} \tag{4}$$

Where $P_i(c)$ denote $P(c \mid x_i)$ and λ is the regularization parameter. According to optimize results of EM likelihood estimation, we get the three parameters Φ_k , μ_k and covariance Σ_k of our Gaussian mixture model.

$$\Phi_k = \frac{1}{m}\sum_{i=1}^{m} p(c_k \mid x_i) \tag{5}$$

$$\mu_k = x_i - \frac{\lambda \sum_{i,j=1}^{m} \{(x_i - x_j)(P(c_k \mid x_i) - P(c_k \mid x_j))\} W_{ij}}{2\sum_{i=1}^{m} P(c_k \mid x_i)} \tag{6}$$

$$\Sigma_k = \sum_{i=1}^{m} \varphi_{i,k} + \frac{\lambda \sum_{i,j=1}^{m} \{[\varphi_{j,k} - \varphi_{i,k}](P(c_k \mid x_i) - P(c_k \mid x_j))\} W_{ij}}{2\sum_{i=1}^{m} P(c_k \mid x_i)} \tag{7}$$

According to the formula derived (5-7), the parameters Φ_k , μ_k and Σ_k will be re-estimated in the objective function. Then the E-Step and M-Step are alternated until a termination condition is met.

IV. EXPERIMENTS

A. Experiment Datasets

To test the algorithmic performance, we use two benchmark datasets of MovieLens data. The first version of MovieLens dataset (ML.A) consists of 943 users, 1682 movies, and 100000 ratings with five level ratings from 1 (i.e., worst) to 5 (i.e., best), additional, each user has rated at least 20 movies. The second version of MovieLens dataset (ML.B) consists of 6040 users, 3900 movies, and 1000209 ratings; also each user has rated at least 20 movies. ML.B dataset includes three files: movies.dat, user.dat and ratings.dat, which recorded movie information, demographic characteristics of users and user ratings on each movie. The detailed description of ML.A and ML.B dataset is shown as TABLE.I and TABLE.II.

TABLE I. THE DESCRIPTION OF ML.A DATASET

ML.A dataset	The description to dataset
u.data	user id, item id, rating, timestamp
u.info	The number of users, items, ratings
u.item	Information about the items (movies),contain: movie id, movie title, release date, et al.
u.genre	a list of the genres
u.user	demographic information about the users, contain: user id, age, gender, occupation, zip code
u.occupation	a list of the occupations.

TABLE II. THE DESCRIPTION OF ML.B DATASET

ML.B dataset	The description to dataset
user.dat	user id, gender, age, occupation, zip-code
movie.dat	movie id, title, genres
Ratings.dat	user id, movie id, rating, timestamp

B. Experiment Results and Analysis

Summaries of the experiment results for comparing algorithms on the two version MovieLens datasets are shown respectively in Fig.2 and Fig.3. We adjust the number of most similar k , to get different precision and recall of the *TopN* recommended in our experiment.

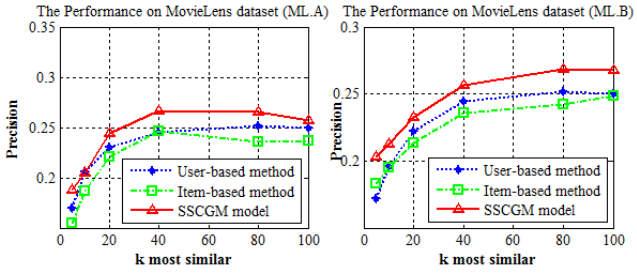

Fig.2. Comparison of percision among User-based CF method, Item-based CF method and SSCGM model method

Fig.2 illustrates the precision of *TopN* recommended, while comparing algorithm performs with different most similar k . Clearly, the recommended prediction based on SSCGM model outperforms User-based CF methods and Item-based CF methods. Among all three comparing algorithms, our algorithm gets the highest precision on the first version of MovieLens dataset (ML.A) when similar $k > 10$, and it gets also the highest precision on the second version of MovieLens dataset (ML.B) with different most similar k . Especially, comparing with User-based CF methods, they get nearly the same recommended precision when k less than or equal 10, but our algorithm increases by about 3 percent when k exceeds 20 on ML.A dataset . We analysis it may be that our algorithm not only to consider user's behavior information but also consider user's demographic characteristics in the computation of the user similarity, which can lead to acquire the similar users more accurate. From the right subgraph of Fig.2, we can get the similar conclusion on ML.B dataset that our method can increases by about 3 percent with different k .

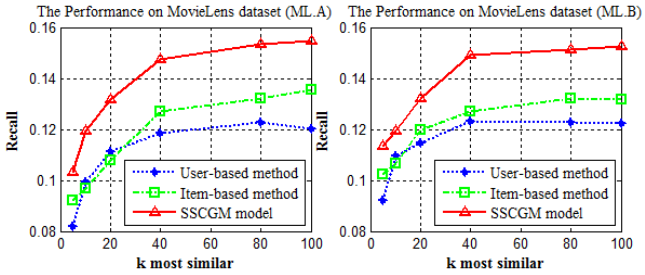

Fig.3. Comparison of Recall among User-based CF method, Item-based CF method and SSCGM model method

Fig.3 illustrates the Recall of *TopN* recommended by comparing algorithm performs with different most similar k . Clearly, our algorithm gets the highest recall with different most similar k on ML.A dataset and ML.B dataset. The result

shows that our proposed SSCGM model method can effectively improve about 2 percent when $k = 5$ in recall, and it can get more improvements when k exceeds 10 on ML.A dataset and ML.B dataset.

From experiment data of Fig.2 and Fig.3, which shows that our algorithm has acquired relatively stable performance, which may be due to considering these two factors that user's behavior information and user's demographic characteristics in the computation of the user's similarity. This confirm the final idea that construct hybrid recommender system using semi-supervised clustering based on Gaussian mixture model, which can appropriately utilizing the advantage of these cluster method to enhance the recommendation performance by utilizing different particular type data, and constructing a hybrid recommendation method based on demographic method and user-based CF method.

V. CONCLUSIONS

This paper describes a novel personalized recommendation method that utilizes semi-supervised clustering based Gaussian mixture model, which provides a hybrid recommender method by combining demographic method and user-based CF method. Our method provides the recommendations for the target user with good quality rating using similarity measures of comprehensive factors. The result from various simulations using MovieLens data set shows that the proposed recommender method performs better than user-based CF method and item-based CF method, which helps to improve the quality of rating.

Focusing on the recommender process, it is powerless that our recommender method in solving the cold-start problem and the ranking should be computed in real time has been considered. In the future, we plan to study problems such as how our system can communicate its reasoning to users; the minimum amount of data (ratings or textual information) required to return accurate recommendations, and a more elaborate way of including item information.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their suggestions to significantly improve this paper. The paper has been partially supported by Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No.KJ1500920 and No.KJ1500926), Innovation Team of Chongqing University of Technology (Grant No.2015TD14).

REFERENCES

[1] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(6): 734-749.

[2] Shinde S K, Kulkarni U. Hybrid personalized recommender system using centering-bunching based clustering algorithm [J]. Expert Systems with Applications, 2012, 39(1): 1381-1387.

[3] Chen L, de Gemmis M, Felfernig A, et al. Human decision making and recommender systems [J]. ACM Transactions on Interactive Intelligent Systems, 2013, 3(3): 17.

[4] Ghazarian S, Nematbakhsh M A. Enhancing memory-based collaborative filtering for group recommender systems [J]. Expert Systems with Applications, 2015, 42(7): 3801-3812.

[5] Pujahari A, Padmanabhan V. An Approach to Content Based Recommender Systems Using Decision List Based Classification with k-DNF Rule Set[C]// Proceedings of 2014 International Conference on Information Technology (ICIT), IEEE, 2014: 260-263.

[6] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey [J]. Knowledge-Based Systems, 2013, 46: 109-132.

[7] De Campos L M, Fernández-Luna J M, Huete J F, et al. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks[J]. International Journal of Approximate Reasoning, 2010, 51(7): 785-799.

[8] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 287-296.

[9] Liu J, Wu C, Liu W. Bayesian probabilistic matrix factorization with social relations and item contents for recommendation [J]. Decision Support Systems, 2013, 55(3): 838-850.

[10] Vozalis M G, Margaritis K G. Using SVD and demographic data for the enhancement of generalized collaborative filtering [J]. Information Sciences, 2007, 177(15): 3017-3037.

[11] Porcel C, Tejeda-Lorente A, Martínez M A, et al. A hybrid recommender system for the selective dissemination of research resources in a technology transfer office [J]. Information Sciences, 2012, 184(1): 1-19.

[12] Choi K, Yoo D, Kim G, et al. A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis [J]. Electronic Commerce Research and Applications, 2012, 11(4): 309-317.

[13] Gao L, Li C. Hybrid personalized recommended model based on genetic algorithm[C] //Proceedings of 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM'08. IEEE, 2008: 1-4.

[14] Verma A, Virk H K. A Hybrid Genre-based Recommender System for Movies using Genetic Algorithm and KNN Approach [J]. International Journal of Innovations in Engineering and Technology, 2015, 5(4): 48-55.

[15] Gupta A, Tripathy B K. A generic hybrid recommender system based on neural networks[C]// Proceedings of 2014 IEEE International Advance Computing Conference (IACC). IEEE, 2014: 1248-1252.

[16] De Campos L M, Fernández-Luna J M, Huete J F, et al. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks[J]. International Journal of Approximate Reasoning, 2010, 51(7): 785-799.

[17] Shah J M, Sahu L. A Hybrid Based Recommendation System based on Clustering and Association [J]. Binary Journal of Data Mining & Networking, 2015, 5(1): 36-40.