Journal of Southeast University (English Edition) Vol. 31, No. 4, pp. 462 - 468 Dec. 2015 ISSN 1003—7985

A novel similarity measurement approach considering intrinsic user groups in collaborative filtering

Gu Liang Yang Peng Dong Yongqiang

(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)
(Key Laboratory of Computer Network and Information Integration of Ministry of Education, Southeast University, Nanjing 211189, China)

Abstract: To improve the similarity measurement between users, a similarity measurement approach incorporating clusters of intrinsic user groups (SMCUG) is proposed considering the social information of users. The approach constructs the taxonomy trees for each categorical attribute of users. Based on the taxonomy trees, the distance between numerical and categorical attributes is computed in a unified framework via a proper weight. Then, using the proposed distance method, the naïve k-means cluster method is modified to compute the intrinsic user groups. Finally, the user group information is incorporated to improve the performance of traditional similarity measurement. A series of experiments are performed on a real world dataset, MovieLens. Results demonstrate that the proposed approach considerably outperforms the traditional approaches in the prediction accuracy in collaborative filtering.

Key words: similarity; user group; cluster; collaborative filtering

doi: 10.3969/j.issn.1003-7985.2015.04.006

As the current innovations in the information and Internet technology boom, people are facing the problem of information overload. The significance of recommendations becomes heightened due to people's inability to find the most interesting and valuable information on the Internet. The research of the recommendation system is ongoing in different areas, e.g., e-commerce[1], social networks[2] and the TV system[3]. Generally speaking, recommendation systems consist of three prevalent methods, the content-based method, collaborative filtering (CF) and sequential pattern analysis. Among these methods, collaborative filtering, first proposed by Goldberg et al. in 1992[4], has been widely studied and applied due to its effectiveness and simplicity.

Generally speaking, the model-based methods and memory-based methods are the main CF techniques[5-6]. The memory-based methods perform better than the model-based methods in some aspects and thus attract considerable attention in this research area. Given an unknown rating on a test item from a test user, the memory-based CF measures the similarity between the test user and other users (user-based) or the similarity between the test item and other items (item-based). Then, the rating to be predicted can be computed by averaging the weighted previous ratings on the test item from similar users (user-based) or by averaging the weighted previous ratings on the similar item from the test user (item-based).

As we can see that, the similarity measurement is a fundamental step in both user-based and item-based methods. Researchers have put forward quite a few similarity measurement methods, including the cosine-based method (COS), Pearson correlation coefficient (PCC) and Euclidean distance (ED)[7-9]. In particular, the COS focuses more on the angle between the vectors to be computed while paying little attention to their lengths. In addition, PCC is used to compare the changing trend of the vector while ignoring the numerical magnitudes. Different from these two approaches, although ED is almost the most traditional method in distance computing, it tends to provide low accuracy due to its simplicity. That is to say, all of them have some inherent defects. Ref. [10] proposed a mitigation method to select different neighbors for each test item. Ref. [10] combined these three methods and provided nine combinations. Besides, a similarity measurement, named Jaccard uniform operator distance, was proposed in Ref. [11] to effectively measure the similarity aiming at unifying similarity comparison for vectors in different multidimensional vector spaces and handling dimension-number difference for different vector spaces. Different from Ref. [11], Ref. [12] argued that traditional similarity measures can be improved by taking into account the contextual information drawn from users. An entropy-based neighbor selection approach for collaborative filtering was put forward in Ref. [13]. The proposed

Received 2015-03-22.

Biographies: Gu Liang (1989—), male, graduate; Yang Peng (corresponding author), male, doctor, associate professor, pengyang@seu.edu.cn.

Foundation items: The National High Technology Research and Development Program of China (863 Program) (No. 2013AA013503), the National Natural Science Foundation of China (No. 61472080, 61370206, 61300200), the Consulting Project of Chinese Academy of Engineering (No. 2015-XY-04), the Foundation of Collaborative Innovation Center of Novel Software Technology and Industrialization.

Citation: Gu Liang, Yang Peng, Dong Yongqiang. A novel similarity measurement approach considering intrinsic user groups in collaborative filtering[J]. Journal of Southeast University (English Edition), 2015, 31(4): 462 - 468. [doi: 10.3969/j.issn.1003-7985.2015.04.006]

method incorporates similarities and uncertainty values to solve the optimization problem of gathering the most similar entities with minimum entropy difference within a neighborhood. Although some of these methods mentioned above improve the recommendation accuracy to some extent, they do not make full use of social information. Some research results on semantic information have also been presented in recent years. Ref. [14] put forward a clustering approach for categorical data based on Tax-Map. Ref. [15] proposed a probabilistic correlation-based similarity measure to enrich the information of records, by considering correlations of tokens. A semantic measure named link weight was demonstrated in Ref. [16], in which the semantic characteristics of two entities and Google page count are used to calculate an information distance similarity between them. The above works make some achievements in similarity measurement while overlooking the significance of numerical data which is considered in this paper. Besides, other neighbor selection approaches were also proposed to improve recommendation quality[17-20].

In this paper, we first propose a novel distance measurement for user record considering its numerical attributes, categorical attributes and the correlation between them. To make the distance metric more reliable, we weigh the attributes by a controlling parameter. Specifically, for the categorical attribute, we build a weighted taxonomy tree to compute the distance. Based on the novel distance measurement, we then attempt to discover the clusters of intrinsic user groups before the similarity computing, i.e., find the neighbors of the test user according to the social information of users. Finally, we propose an incorporation method to compute the similarity between users considering the groups they belong to. The experiments show the advantages of our novel approach over prediction accuracy.

1 Preliminaries

1.1 User-based collaborative filtering

As mentioned above, the memory-based CF method can be divided into user-based and item-based approaches. The recommendation relies on a user-item matrix. This matrix contains the information of users, items and users' ratings. A row vector in the matrix represents a user's ratings on all items, while a column vector expresses the ratings on an item from all users. Note that, the element in the matrix remains null when the item has not been rated by the corresponding user.

Here, we focus on the user-based collaborative filtering. The user-based methods compute the similarity between the test user and others based on their previous ratings on all items. According to the user-item matrix, we can use the three traditional approaches to compute the user similarity. Here, we take the PCC approach as an example. The formulation is as follows:

$$s(A, B) = \frac{\sum_{i \in I_A \cap I_B} (r_{Ai} - \bar{r}_A)(r_{Bi} - \bar{r}_B)}{\sqrt{\sum_{i \in I_A \cap I_B} (r_{Ai} - \bar{r}_A)^2} \sqrt{\sum_{i \in I_A \cap I_B} (r_{Bi} - \bar{r}_B)^2}} \quad (1)$$

where $\bar{r}_A$ and $\bar{r}_B$, respectively, represent the average ratings of users A and B on all the items they have rated; $I_A \cap I_B$ denotes the intersection of the items that users A and B have rated. When $I_A \cap I_B = 1$, $S(A, B)$ equals zero.

After that, the user-based CF sorts the users according to their similarity with the test user. The rating to be predicted is computed by aggregating the ratings from other users with proper weight. The more similar a user is to the test user, the higher the weight assigned to the prediction rating. The detailed aggregating strategy is as

$$r_{Am} = \bar{r}_A + \frac{\sum_{u \in U_A} s(A, u)(r_{um} - \bar{r}_u)}{\sum_{u \in U_A} s(A, u)} \quad (2)$$

where U_A is the set of users similar to user A; $s(A, u)$ is computed according to Eq. (1). In particular, r_{Am} is equal to the average rating of user A when there are no similar users for him.

1.2 k-means clustering

In data mining area, k-means clustering is a well-known method for cluster analysis aiming to partition n observations into k clusters, in which each observation belongs to the cluster with the nearest mean. The rationale of k-means clustering can be illustrated as follows: Given a set of observations $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$, where each observation is a multi-dimensional real vector, k-means clustering attempts to partition the n observations into $k(\leqslant n)$ sets $C = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_k\}$ so as to minimize the within-cluster sum of squares. In other words, its objective function is

$$C = \arg\min_{C} \sum_{i=1}^{k} \sum_{\boldsymbol{X}_j \in \boldsymbol{C}_i} \| \boldsymbol{X}_j - \overline{\boldsymbol{C}_i} \| \quad (3)$$

The k-means clustering technique has been proved to be useful in many applications. Notice that, k-means clustering cannot deal well with categorical attributes due to its distance metric in clustering iterations.

2 A Novel Similarity Measurement Approach

In this section, we describe our proposed approach in detail. First, we give a new definition of the distance metric in clustering aiming to deal with numerical and categorical attributes in a unified model. Then, we present the clustering process of discovering the intrinsic user groups. Finally, we show the proposed similarity measurement approach based on the intrinsic user groups.

2.1 New definition of the distance metric

The distance function is a critical element in the cluste-

ring problem. Generally speaking, the distance function computes the dissimilarities among data points (two-dimensional) or hyper-points (n-dimensional, $n > 2$). Choosing an appropriate distance metric is important for obtaining an accurate result under attributes of specific types (numerical or categorical) or different sizes.

Unlike the normal attributes in the clustering problem, the attributes in CF technique typically consist of both the numerical and categorical attributes and every attribute always has a unique scale. Hence, in the CF area, we need a new distance metric to handle the above features. Ref. [21] introduced a measure that uses the simple matching similarity measure for categorical attributes. However, the measure in Ref. [21] cannot deal well with the attributes of user information in CF due to its indiscrimination of the distance between different categorical elements in the same attribute.

In this paper, we propose a new definition of the distance metric by considering the normalization of both the numerical and categorical attributes and the effect of the association-rule-based taxonomic tree. Here, we provide the definitions of numerical distance and categorical distance including the normalization.

Definition 1 (numerical distance) Let $n_{\min}$ and $n_{\max}$ be the minimum and maximum values of a numerical attribute. Given that two values n_1 and n_2 belong to this numerical attribute, the normalized distance is defined as

$$N(n_1, n_2) = \frac{|n_1 - n_2|}{|n_{\min} - n_{\max}|} \tag{4}$$

We can take a typical numerical attribute Age as an example to illustrate the numerical distance further. Consider the records in Tab. 1. The distance contributed by the Age attribute for the first two records is $|12-24|/|12-53| = 0.292$, while the distance between the first and the third records with respect to the same attribute is $|12-40|/|12-53| = 0.682$. The smaller the distance between two values, the more similar they are. Clearly, the first record is more similar to the second record than the third one.

Tab. 1 Typical cases

Age	Gender	Occupation	Salary	Zipcode
12	M	Student	0	27510
24	M	Artist	30 000	10003
40	F	Librarian	16 000	30030
49	M	Engineer	50 000	55107
53	F	Lawyer	60 000	90703
38	F	Engineer	38 000	48197
29	M	Lawyer	42 000	55369

As the categorical attributes cannot be converted into numerical values, it is difficult to compute the distance between two values under some categorical attribute directly. One solution is that, if the two values under the attribute are the same to each other, the distance between them is 0. Otherwise, the distance is 1. Besides, Ref. [21] captured the semantic relationship among the values and built the taxonomy tree for them, thus improving the distance accuracy to some extent. However, this method faces difficulty when the two values belong to the same level of the taxonomy tree. In this paper, we attempt to solve this problem by discovering their association rules with other numerical attributes.

Definition 2 (categorical distance) Let $V = \{C_1, \ldots, C_p, \ldots, N_1, \ldots, N_q\}$ be a record including p categorical attributes $\{C_1, C_2, \ldots, C_p\}$ and q numerical attributes $\{N_1, N_2, \ldots, N_q\}$. Let $T_h (h \in [1, q])$ be a taxonomy tree for C_h. y_i, y_j are two values from the same categorical attribute C_h, and $N_s (s \in [1, p])$ is a numerical attribute that has a value interval $[n_{\min}, n_{\max}]$. The normalized distance between y_i and y_j is defined as

$$C(y_i, y_j) = \frac{H(T(y_i, y_j))}{H(T_h)} \times N\mathrm{dis}(\overline{n_{s, y_i}}, \overline{n_{s, y_j}}) \tag{5}$$

where $T(y_i, y_j)$ is the subtree rooted at the lowest common ancestor of y_i and y_j; $H(T_h)$ represents the height of the tree h; $\overline{n_{s, y_i}}$ and $\overline{n_{s, y_j}}$, respectively, denote the average value of N_s in all the records appearing simultaneously with y_i and y_j. Specially, taking y_i as an example, the formulation is as

$$\overline{n_{s, y_i}} = \frac{\sum_{t=1}^{N} n_{y_i}}{N} \tag{6}$$

where N is the number of all the records.

A simple case is shown in Fig. 1. Fig. 1 illustrates the taxonomy tree of the attribute Occupation in Tab. 1. In this case, every profession is equal in the taxonomy tree and the distance between them is 0 without considering the association rules with other numerical attributes like Salary. However, it is not difficult to infer that any profession should have some underlying correlation with other professions. This paper attempts to discover this correlation. With the function proposed in Definition 2, we can discover the association rule between Occupation and Salary. The new distance between the attribute Occupation of records 4 and 5 is

$$\frac{1}{1} \times \frac{(50\,000 + 38\,000)/2 - (60\,000 + 42\,000)/2}{60\,000 - 0} = 0.117$$

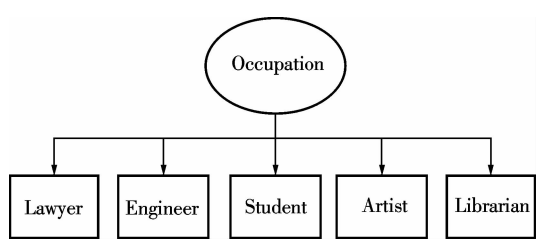

Fig. 1 Taxonomy tree of Occupation

How to construct the taxonomy tree of each attribute is a key point in our approach. Generally speaking, re-

searchers construct the tree manually according to the domain knowledge or use the decision tree algorithm, e. g., ID3 and C4. 5. The former possesses better performance than the latter while having worse operability when the attributes are complicated. In our proposed approach, we construct the taxonomy tree manually to obtain better performance considering that the user attribute in this paper is relatively simple.

Definition 3 (record distance) Given two records r_1 and r_2 with the attributes as introduced in Definition 2, the distance between them is defined as

$$R(r_1, r_2) = \lambda \sum_{i=1}^{p} C(r_1[C_i], r_2[C_i]) + (1 - \lambda) \sum_{j=1}^{q} N(r_1[N_j], r_2[N_j]) \quad (7)$$

where $r_i[x]$ represents the value of attribute x in r_i; C and N are defined in Definitions 1 and 2, respectively; C_x is the center of the cluster which the record x belongs to; and λ is a weight parameter to control the contributions of numerical attributes and categorical attributes. Notice that, when λ is equal to 0, the distance between the records is entirely dependent on their numerical attributes and this can deal well with the cases that user records have few or no categorical attributes.

2.2 Discovering intrinsic user groups

Based on the distance metric proposed above, in this part, we attempt to discover the intrinsic user groups using the k-means clustering technique. We first give the definition of intrinsic user groups in our approach.

Definition 4 (intrinsic user groups) Given a user record set U, it will be divided into m intrinsic user groups, $\{g_1, g_2, \ldots, g_m\}$ according to the record distance defined in Definition 3 so that, for each user record u in U, if u is grouped into g_i, two conditions must be satisfied:

$$\min \sum_{v \in g_i} R(u, v) \quad (8)$$

$$\max \sum_{v \in U, v \notin g_i} R(u, v) \quad (9)$$

The intrinsic user groups can be obtained by the record distance between user records. Given the initial set of records, the k-means algorithm can be divided into three distinct phases: initial, assignment and update phase. In the initial phase, k points are selected as the initial centers of k clusters. In the assignment phase, each point is assigned to the closest center according to a distance metric. While in the update phase, the cluster centers of any changed clusters are recomputed as the average of members of each cluster. The last two phases are executed iteratively until the algorithm converges. We set up the brief process to discover the intrinsic user groups as follows.

Algorithm 1 Discovering algorithm

Input: a positive integer k, an iteration number m, a convergence threshold δ_0, a set of user records S.

Output: a set of k groups and their centers.

```
If ( |S| < k)
  Return;
  End If;
  Pick k user records as centers randomly, cost = MAX;
    While( m > 0 || δ < δ0)
    For i = 1, 2, …, |S|
      For j = 1, 2, …, k
          N(Si, gj);
          C(Si, gj);
          R(Si, gj);
        End For;
        c = Min_Rdis(Si);
        gc ← Si;
      End For;
      δ = |C(g) − cost|;
      cost = Cost(g);
      For i = 0, 1, …, k
      Center(gk);
      End For;
      m = m − 1;
      End While;
      Return gm and Center(gm), m = 1, 2, …, k;
End;
```

In Algorithm 1, $\text{Min\_}R\text{dis}(S_i)$ is the function to obtain the center closest to S_i. $C(g)$ is computed using Eq. (3). Center(g_k) represents the center of g_k. Once Algorithm 1 is finished, we obtain the k intrinsic user groups.

2.3 CF with Novel Similarity Measurement Approach

In Section 2. 2, we have discovered the intrinsic user groups by a new distance metric. Then, we incorporate this information to compute the similarity between users. The incorporation strategy can be illustrated as

$$S(A, B) = \frac{\sum_{i \in I_A \cap I_B} \frac{(r_{Ai} - \bar{r}_A)(r_{Bi} - \bar{r}_B)}{R(C_A, C_B)}}{\sqrt{\sum_{i \in I_A \cap I_B} \frac{(r_{Ai} - \bar{r}_A)^2}{R(C_A, C_B)}} \sqrt{\sum_{i \in I_A \cap I_B} \frac{(r_{Bi} - \bar{r}_B)^2}{R(C_A, C_B)}}} \quad (10)$$

where r_{yi} is the rating of user y on item I; $\bar{r}_y$ represents the average rating of user y; C_x is the center of the cluster or the user group which the record x belongs to; $R(C_A, C_B)$ is the record distance defined in Definition 3. Finally, we can simply make predictions with Eqs. (10) and (2). An obvious modification is that, our proposed approach completely utilizes the rating information and the user social information, rather than purely the former one as in traditional PCC approaches.

3 Empirical Analysis

This section describes the experimental design for evaluating the proposed similarity measurement approach, as well as how the approach affects the quality of recommendation. The implication of the experiments is also introduced in this section.

3.1 Dataset

In order to evaluate the performance of our approach, we perform the experiments on the MovieLens dataset, which is a well-known dataset for collaborative filtering collected by the GroupLens research team at the University of Minnesota. The dataset includes 100 000 ratings on 1 682 items by 943 users. Moreover, the rating scale of the dataset is from 1 to 5 and each user rated at least 20 movies. To obtain reliable experimental results, 90% of each target user's ratings are used as training data, and the remaining ratings are used as test data.

3.2 Evaluation metrics

The accuracy of prediction is the most common assessment criteria in CF area. We use the well-known mean absolute error (MAE) to evaluate the prediction accuracy. MAE is the average absolute deviation of predictions to the ground truth, which is defined as

$$\mathrm{MAE} = \frac{\sum_{u,i} | r_{u,i} - r'_{u,i} |}{N} \tag{11}$$

where $r_{u,i}$ denotes the real rating on item i from user u; $r'_{u,i}$ denotes the predicted rating; and N represents the total number of all the test ratings. The smaller the value of MAE, the better the performance.

3.3 Performance comparison

3.3.1 Comparisons with other traditional approaches

In order to illustrate the effectiveness of our proposed approach SMCUG, we compare it with five representative similarity measurement approaches: COS[6], PCC[7], ED[8], CF_P_D[9] and CBPCC[18]. In particular, Ref. [9] introduces nine combination methods and CF_P_D shows the best performance among them on the Movielens dataset. According to Definition 3, we can observe that λ is a significant parameter. In this experiment, we set λ to be 0.5. That is, the categorical attributes and numerical attributes of user record have equal contributions to the clustering of intrisic user groups. We attempt to group all users into 50 groups by setting the parameter k to be 50 in the clustering process. We vary the neighbor's size from 5, 20, 40, 60, 80, to 100. Fig. 2 shows the MAE performance comparison of all the evaluated approaches. From Fig. 2 we can infer that, as the neighbors number increases, all the approaches tend to obtain lower MAE results, which means more accurate predictions. Among them, the ED approach obtains a relatively high MAE result. We believe that this is caused by its inherent metric limitation. Our proposed approach outperforms all the other approaches with different numbers of neighbors.

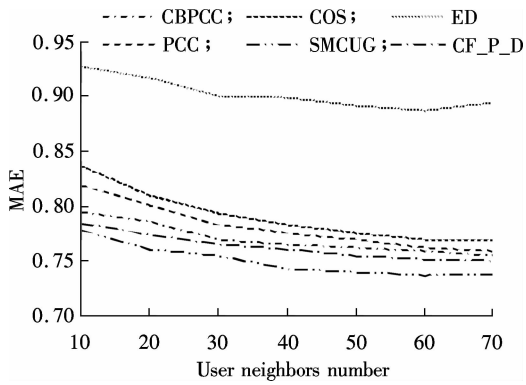

Fig. 2 MAE plots of all the approaches with different numbers of neighbors

3.3.2 Impacts of factors

In our proposed approach, the cluster number and attribute factor have significant effects on the final predictions. We let one of them be a constant and then observe the effect of the other on the prediction result. First, the cluster number parameter k is set to be 50 and we vary the attribute factor λ from 0 to 1. The experimental result is illustrated in Fig. 3(a). As can be seen, we conduct the experiments when the neighbors number is 10, 20, and 40. Under these three conditions, the MAE curves with different neighbor numbers are similar. The most accurate prediction can be obtained around the value of 0.4. We

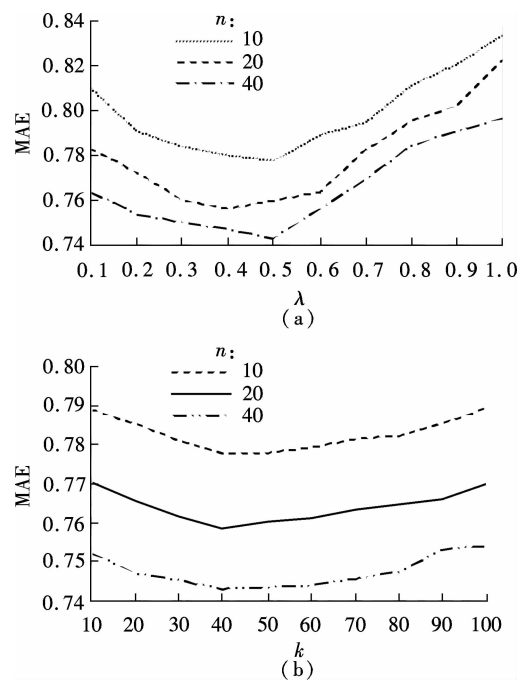

Fig. 3 MAE plots of SMCUG with different λ and k. (a) Plots with different λ ($k=50$); (b) Plots with different k($\lambda=0.5$)

hold that this is mainly because our proposed approach assigns an appropriate weight to both the numerical and categorical attributes at this point. The numerical attributes seem more important for prediction accuracy than the categorical attributes. As for other datasets, we can train the parameter λ with a small part of the dataset to ensure a satisfactory result due to the fact that the dataset feature of one application tends to be stable as its data size increases.

Fig. 3(b) illustrates the effect of user groups number on overall prediction accuracy. The attribute parameter λ is set to be 0.5. From Fig. 3(b), it is apparent that the number of user groups does have an effect on the performance of our approach. As the number of user groups increases, the MAE of our approach descends until the number reaches around 40. After then, the MAE goes up again when the number varies from 40 to 100. We infer that, the large number of user groups makes the user information more specific, thus leading to the overfitting problem. Moreover, the small number of user groups makes the groups imprecise and we cannot utilize the intrinsic information adequately. Both the conditions are detrimental to the prediction accuracy.

4 Conclusion

We propose a novel similarity measurement approach incorporating clusters of intrinsic user groups in collaborative filtering. Due to the proper clustering technique, our approach can utilize the user social information effectively and improve the prediction results notably. Experiments performed on a real-world dataset demonstrate that our proposed approach outperforms other approaches. In the future, we plan to conduct a better analysis of the approach and focus on the item grouping.

References

[1] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews [C]//*Proceedings of the* 1994 *ACM Conference on Computer Supported Cooperative Work*. Chapel Hill, NC, USA, 1994: 175 - 186.

[2] Walter F E, Battiston S, Schweitzer F. A model of a trust-based recommendation system on a social network [J]. *Autonomous Agents and Multi-Agent Systems*, 2008, **16**(1): 57 - 74.

[3] Hsu S H, Wen M H, Lin H C, et al. AIMED—a personalized TV recommendation system[M]//*Interactive TV*: *a shared experience*. Berlin: Springer, 2007: 166 - 174.

[4] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. *Communications of the ACM*, 1992, **35**(12): 61 - 70.

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

[6] Sahoo N, Singh P V, Mukhopadhyay T. A hidden Markov model for collaborative filtering[J/OL]. *Management Information Systems Quarterly*, 2012. http://ssrn.com/abstract=1700585.

[7] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//*Proceedings of the* 10*th International Conference on World Wide Web*. Hong Kong, China, 2001: 285 - 295.

[8] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering [C]//*Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Amsterdam, Holland, 2007: 39 - 46.

[9] Kim H K, Kim J K, Ryu Y U. Personalized recommendation over a customer network for ubiquitous shopping [J]. *IEEE Transactions on Services Computing*, 2009, **2**(2): 140 - 151.

[10] Choi K, Suh Y. A new similarity function for selecting neighbors for each target item in collaborative filtering [J]. *Knowledge-Based Systems*, 2013, **37**(2): 146 - 153.

[11] Sun H F, Chen J L, Yu G, et al. JacUOD: a new similarity measurement for collaborative filtering[J]. *Journal of Computer Science and Technology*, 2012, **27**(6): 1252 - 1260.

[12] Bobadilla J, Ortega F, Hernando A. A collaborative filtering similarity measure based on singularities[J]. *Information Processing & Management*, 2012, **48**(2): 204 - 217.

[13] Kaleli C. An entropy-based neighbor selection approach for collaborative filtering[J]. *Knowledge-Based Systems*, 2014, **56**(3): 273 - 280.

[14] Dos Santos T R L, Zárate L E. Categorical data clustering: what similarity measure to recommend?[J]. *Expert Systems with Applications*, 2015, **42**(3): 1247 - 1260.

[15] Song S, Zhu H, Chen L. Probabilistic correlation-based similarity measure on text records[J]. *Information Sciences*, 2014, **289**(5): 8 - 24.

[16] Jiang Y, Wang X, Zheng H T. A semantic similarity measure based on information distance for ontology alignment[J]. *Information Sciences*, 2014, **278**(10): 76 - 87.

[17] Xue G R, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based smoothing[C]//*Proceedings of the* 28*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Singapore, 2005: 114 - 121.

[18] Roh T H, Oh K J, Han I. The collaborative filtering recommendation based on SOM cluster-indexing CBR[J]. *Expert Systems with Applications*, 2003, **25**(3): 413 - 423.

[19] Honda K, Sugiura N, Ichihashi H, et al. Collaborative filtering using principal component analysis and fuzzy clustering[M]//*Web intelligence*: *research and development*. Berlin: Springer, 2001: 394 - 402.

[20] Bilge A, Polat H. A comparison of clustering-based privacy-preserving collaborative filtering schemes[J]. *Applied Soft Computing*, 2013, **13**(5): 2478 - 2489.

[21] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values[J]. *Data Mining & Knowledge Discovery*, 1998, **2**(3): 283 - 304.

一种协同过滤中考虑潜在用户分组的相似度度量方法

顾 梁 杨 鹏 董永强

(东南大学计算机科学与工程学院,南京 211189)
(东南大学计算机网络和信息集成教育部重点实验室,南京 211189)

摘要:为了提高用户之间相似度度量的性能,充分利用用户的社会信息,提出一种考虑潜在用户分组信息的相似度度量方法. 该方法首先为用户的分类属性建立权值分类树,并基于此分类树,采用统一框架计算用户分类信息和数值信息的距离;然后利用该距离改进 k-means 聚类方法,以计算用户的潜在用户分组;最后结合用户分组信息改进传统相似度度量方法. 基于真实数据集 MovieLens 进行实验,并与其他传统方法对比,结果表明,与传统方法相比,所提方法提高了协同过滤中的预测精度.

关键词:相似性; 用户组; 聚类; 协同过滤
中图分类号:TN92