3M2-IOS-3b-5

A Median Concept for Model-Based Collaborative Filtering

Ekkawut Rojsattarat*1 Pakaket Wattuya*2

*1Department of Computer Science, Kasetsart University Si Racha Campus
*2Department of Computer Science, Kasetsart University

Model-based collaborative filtering has widely been used to allow the recommender systems to learn to recognize complex patterns of user preferences. So far many clustering models have been investigated to solve the shortcomings of memory-based algorithms. Mostly, the user preferences are in form of sparse matrix which comprises of many noises. Predictions of recommendations for any active users are unavoidably effected by outlier data. While the previous works rely on solving general clustering problems, we propose a general framework to apply a concept of generalized median to collaborative filtering. From a general point of view, given a distance function $d(p, q)$, the essential information of a given set of patterns S in arbitrary space U is covered by a generalized median $p \in U$ that minimizes the sum of distances to all patterns from S. This concept has found various applications in dealing with strings, graphs, curves, and clusterings. In our current work pseudo user rating matrix which obtains by item-based collaborative filtering to control a degree of data sparsity. The various model-based collaborative filtering by generalized median concept is applied to compare overall performance in reducing the leftover effects of outliers. Additionally, we compare our approach to a model-based collaborative filtering algorithms using K-mean clustering for performance evaluation.

1. Introduction

Collaborative filtering (CF) is a technique for producing a recommendation to a target user based on implicit or explicit ratings given by a group of users. It has been widely used in many online commercial stores and social communities in various domains of filtered items, such as, films, music, books, Usenet news, etc. CF techniques use preferences for items by users to predict additional items, topics or products a new user might like in form of user-item rating matrix (see Table 1). The rating can either be explicit on predifined range of scale such as, 1-5 scale or implicit indications such as a purchase or click-throughs [Miller 04]. Typically, there are missing values in the matrix where users did not give their preferences for certain items.

Collaborative filtering methods can be categorized into two approaches, memory-based and model-based. Memory-based approache operates on the entire user-item rating matrix and generates recommendations by identifying the neighborhood of the target user to whom the recommendations will be made, based on the agreement of other users past ratings. Model-based techniques use the rating data to train a model and then the model will be used to derive the recommendations. Well-known model-based techniques include clustering [Ungar 98], machine learning on the graph [Zhou 08], matrix factorization (e.g. SVD) [Koren 09], etc. So far memory-based techniques are quite successful in real-world applications because they are easy to understand, easy to implement and work well in many real-world situations. However, there are some problems which limit the application of memory-based techniques, especially in the large-scale applications. The most serious problem is the sparsity of user-item rating matrix where each user only rates a small set of items. The similarity between users (or items) is often derived from few overlapping ratings and it is hence a noisy and unreliable value. Another problem of memory-based CF is efficiency. It has to compute the similarity between every pair of users (or items) to determine their neighborhoods. This is not computationally feasible for the online systems with millions of users and items. to overcome the weakness of memory-based CF, researchers have focused on model-based clustering techniques with the aim of seeking more accurate, yet more efficient methods. Based on ratings, these techniques group users or items into clusters, thus give a new way to identify the neighborhood.

Contact: Ekkawut Rojsattarat, Department of Computer Science, Kasetsart University Si Racha Campus, 199 Thungsukla, Si Racha, Chonburi, Thailand, +66 38354587-8, frenviekr@src.ku.ac.th

Table 1: User-Item Rating Matrix.

	Item 1	Item 2	...	Item n
User 1	$r_{1,1}$	$r_{1,2}$	...	$r_{1,n}$
User 2	$r_{2,1}$	$r_{2,2}$	...	$r_{2,n}$
...	...	....	...	...
User m	$r_{m,1}$	$r_{m,2}$	...	$r_{m,n}$

In this work, we focus on a model-based CF based on applying a generalized median concept to clustering methods. The main idea of this work is instead of using K-mean algorithm to define a center of each cluster, a generalized median vector produced by Weiszfeld algorithm [Weiszfeld 37] has been used to define a representation of each partition. Additionally, to overcome the problem of matrix sparsity a pseudo rating matrix generated by Slope one algorithm [Lemire 05] has been investigated.

The rest of this paper is organized as follows. We review current cluster methods on collaborative filtering in section 2, and propose our methods in section 3. In Section 4 we report some experimental results. And finally, some

discussions conclude this paper.

2. Related Work

So far various clustering technique was applied to collaborative filtering researchs. In [Ungar 98], they proposed clustering method for collaborative filtering. they clustered users and items separately. With the clustering methods, they reduced the sparse problem but failed to improve the accuracy. OConner, M. et al. proposed collaborative based on item clustering [OConner 01]. they reduced one large-dimensionality space into a set of smaller-dimensionality space. Although the scalability was improved, but failed to improve the accuracy. [Xue 05] proposed a cluster-based smoothing method. On the sparse dataset, not only the scability was improved but also the accuracy was improved. In [Ding 05], presented another clustering application. they applied different parameter of time function on users in different clusters and the accuracy was improved by the differentiation of the parameter.

In the above methods, except in [OConner 01], they applied K-mean clustering method and did not analyze the effect of different clustering methods. moreover, a distance function to measure similarity for cluster was chosen based on the Pearson correlation coefficient.

3. Proposed Method

In this section, we provide some related background knowledges and decribe our model-based collaborative filtering framework in more detail.

3.1 A Generalized Median Concept and The Averaging Problem

From a general point of view the averaging problem can be stated as follows. Assume that we are given a set S of patterns in an arbitrary representation space U and a distance function $d(p, q)$ to measure the dissimilarity between any two patterns $p, q \in U$. The essential information of the given set of patterns is covered by a pattern $p \in U$ that minimizes the sum of distances to all patterns from S, i.e.

$$\bar{p} = \arg\min_{p \in U} \sum_{p \in S} d(p, q)$$

The pattern p is called the *generalized median* of S. If the search is constrained to the given set S, the resultant pattern

$$\hat{p} = \arg\min_{p \in S} \sum_{p \in S} d(p, q)$$

is called the *set median* of S. In general there is no unique solution for both set and generalized median. This concept has found various applications in dealing with strings [Jiang 03], graphs [Jiang 01], curves [Jiang 00], and 2D shapes [Rojsattarat 08].

3.2 The Weiszfeld Algorithm

Since the problem of searching for a generalized median in representation space U is NP-complete problem, we here introduce an iteratively re-weighted least squares method, called Weiszfeld algorithm. The algorithm defines a set of weights that are inversely proportional to the distances from the current estimate to the samples, and creates a new estimate that is the weighted average of the samples according to these weights. That is,

$$\bar{p}_{i+1} = \left(\sum_{p \in S} \frac{p}{d(p - \bar{p}_i)} \right) \Big/ \left(\sum_{p \in S} \frac{1}{d(p - \bar{p}_i)} \right)$$

where $\bar{p}_i$ is a current estimation of generalized median vector, $\bar{p}_{i+1}$ is a new estimation of generalized median vector, and a distance function $d()$ define by the Euclidean distance. This algorithm iteratively run to find a new estimate until a set of weights remain unchange.

3.3 K-Median Clustering for Collaborative Filtering

As we mention in section 1, typically user-rating matrix trend to be a sparse matrix comprise of many missing values. So given a matrix of explicit user ratings, we firstly contruct a pseudo use-rating matrix by using weighted Slope one algorithm [Lemire 05]. The missing ratings are filled in this pseudo user-rating matrix. Then entire users are partitioned into a predefined number of groups by the k-median clustering based on the Weiszfeld algorithm. To find a neighborhood of the target user, instead of looking for a set of users from entire space we only calculate similarities between a target user and the center of each cluster. It can reduce the execution time significantly for the task. We now know that the neighborhood of the target user are users which have been partitioned into the same most similar cluster center. A list of recommendation items is produced by a weighted average as follows.

$$\hat{p}_{t,i} = \bar{r_a} + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u) \times P_{t,u}}{\sum_{u=1}^{n} P_{t,u}}$$

where $\hat{p}_{t,i}$ is the prediction for item i of the target user t; $P_{t,u}$ is the similarity of between the target user t and user u based on the Pearson correlation between their rating vector; and n is a number of neighborhood selected by k-median clustering.

4. Experiments

In this section we present our experimental results of applying our method to generate pridiction. Our results are divide into a series of experiments by two dataset. The aim of our experiments are comparing the performance between our k-median methods based on Weiszfeld algorithm and original k-mean clustering.

4.1 The Dataset

We use two different datasets, MovieLens dataset which provided by the GroupLens Research Project at the University of Minnesota. It consists of 1,000,000 ratings (range from 1-5) from 6040 users and 3500 movies. And another dataset is a sub set of Jester dataset which comprises of

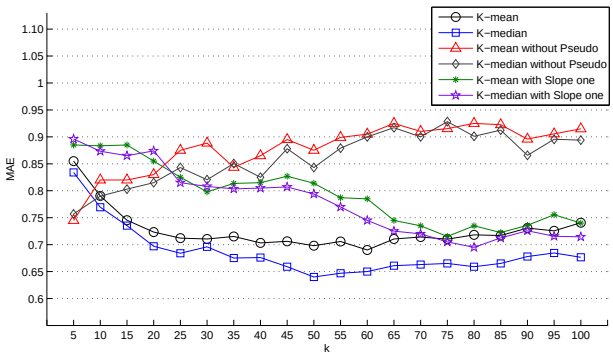

Figure 1: The mean absolute error with different k on MovieLens Dataset

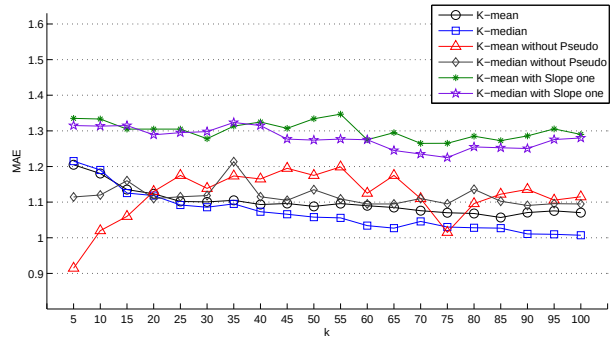

Figure 2: The mean absolute error with different k on Jester Dataset

rating from 24,983 users who have rated 36 or more from total number of 100 jokes.

We conduct experiments using 5-fold cross-validation by ramdomly partition users into 5 groups and retain the training to test ratio at level of 4:1. For performance evaluation, we remove some of the actual ratings within the test set and the mean absolute error (MAE) between the predicted rate and the actual rating of users is calculated. The final results of each experiment are an average over 5 runs.

4.2 Experimental Results

We investigate the overall performance of our approach compare to original k-mean clustering by increasing the value of parameter k from 5 to 100 increment by 5. Two clustering methods are performed on pseudo user-rating matrix of training data. Then we use the correlation-based with neighborhood size $n = 20$ to produce predicted ratings of test data.

We run 2 additional experiments to confirm that our approach can reduced the problem of sparsity and robust when use a different prediction algorithm. For the sparsity problem, K-median and k-mean clustering methods are perform on the original user-rating matrix while other parameter remains the same. Lastly, we conduct an experiment by perform two clustering methods on pseudo user-rating matrix. But we modify a prediction method of the algorithm to Slope one. The results of the experiments on MovieLens and Jester Dataset shows in Figure 1 and Figure 2 respectively.

The experimental results show that our proposed method can significantly reduce the mean absolute error compare to clustering based on k-mean on both datasets. When increasing the value of parameter k, the performance of k-median clustering approach improve quite impressive. The smaller MAE on MovieLens dataset did not tell that our approach can work better than larger MAE on jester dataset. Because they have different rating scales. So far, we can imply that the generalized median not only improve the efficient of running time, but also improve the accuracy of rating prediction. The results of two additional experiments are clear tell that without the help of pseudo user-rating matrix, the accuracy degrade at most level of parameter k. And the last experiment shows that the algorithm trend to still produce a better result when using a different prediction method.

5. Conclusion

Model-based collaborative based on clustering has been widely used in producing a recommendation. In contrast to the existing algorithms which mostly rely on k-mean clustering algorithm, we propose a general framework to apply a generalized median concept to collaborative filtering. an Experimental results have been shown to illustrate the effectiveness of our approach for generating more precie prediction ratings. In real world problem, our model-based approach allows a recommender system can run instantly online. Because clustering algorithms based on Weiszfeld algorithm reduce a large search space to a predefine number of clusters. We have introduced a pseudo user-rating matrix based on slope one technique to reduce the problem in sparse rating matrix. Moreover, experimental results also show that our general framework can be a baseline algorithm to built a better recommender system.

References

[Ding 05] Ding, Y. and Li, X.: Time Weight Collaborative Filtering, Ph.D. Dissertation (2005)

[Jiang 00] Jiang, X., Schiffmann, L. and Bunke, H.: Computation of Median Shapes, In Proceeding of 4th Asian Conference on Computer Vision, pp. 300-305, Taipei (2000)

[Jiang 01] Jiang, X., Munger, A. and Bunke, H.: On Median Graphs: Properties, Algorithms, and Applications, IEEE Transaction on Pattern Analysis and Machine Intelligence, 23(10), pp. 1144-1151 (2001)

[Jiang 03] Jiang, X., Abegglen, K., Bunke, H. and Csirik, J.: Dynamic Computation of Generalized Median Strings, Pattern Analysis and Applications, 6(3), pp. 185-193, (2003)

[Koren 09] Koren, Y., Bell, R. and Valinsky, C.: Matrix Factorization Techniques for Recommender Systems, Computer, 42(8), pp. 30-37 (2009)

[Lemire 05] Lemire, D. and Maclachlan, A.: Slope One Predictors for Online Rating-Based Collaborative Filtering, In Proceedings of SIAM Data Mining (2005)

[Miller 04] Miller, B. N., Konstan, J. A. and Riedl, J.: PocketLens:Towards a Personal Recommender System, ACM Transactions on Information System, Vol.22, No.3, pp. 133-151 (2004)

[OConner 01] OConner, M. and Herlocker, J.: Clustering Items for Collaborative Filtering, In Proceedings of the Workshop on Recommendation Systems, New Orleans, LA (2001)

[Rojsattarat 08] Rojsattarat, E. and Jiang, X.: Shape Averaging by Invariant Fourier Descriptors,In Proceeding of 8th Industrial Conference Advances in Data Mining, pp. 145-153 (2008)

[Ungar 98] Ungar, L. and Foster, D.: Clustering Methods for Collaborative Filtering, In Proceedings of the Workshop on Recommendation Systems, AAAI Press, Menlo Park California. (1998)

[Weiszfeld 37] Weiszfeld, E.: Sur le point pour lequel la somme des distances de n points donns est minimum, Thoku Mathematical Journal, 43, pp. 355386. (1937)

[Xue 05] Xue, G, Lin, C., Yang, Q., Xi, W., Zeng, H., Yu, Y and Chen, Z.: Scalable Collaborative Filtering Using Cluster-Based Smoothing, In Proceeding of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.114-121, New York (2005)

[Zhou 08] Zhou, D., Zhu, s., Yu, K., Song, X., Tseng, B. L., Zha, H. and Giles, C. L.: Learning Multiple Graphs for Document Recommendation, In Proceeding of the 17th International Conference on World Wide Web, pp. 30-37, New York (2008)