Improve recommender systems using certain formulas and fuzzy concept networks

Elaheh Kafshi Tagiabadi
Department of Computer, Islamic Azad University, Mashhad, Iran
e_taghiabadi@mshdiau.ac.ir

Mehrdad Jalali
Department of Computer, Islamic Azad University, Mashhad, Iran
mehrdadjalali@ieee.org

Abstract—**Personalization systems are presented to users to improve suggestion when users are surfing on the net. In order to create accurate model from user webpage contents can be used. Recently some researches were done for adding semantic to user behaviors. However, all users don't show all their interest, but the system must scrutinize user behavior and suggest best prediction for his interest. The best result of this paper is presents a mechanism to improve both user behaviors in web site and exploits fuzzy concept for building automatic model from user interest automatically. Also when the system proposed an item to users in order to calculate item similarity, use collaborative filtering in which to predict user interests uses enriched prediction formula. The results show that the proposed method provides better prediction than traditional methods.**

Keywords-component; trust, recommender systems, fuzzy concept networks, personalization web, collaborative filtering

I. INTRODUCTION

With massive growth in Internet, recommender systems are proposed as solution problems of redundancy information. These systems help users to find the most relevant information that they need. One of the most successful technologies in recommender systems are collaborative filtering systems. These technologies are applied to provide best recommend to customers in commercial systems. The main goal in the collective filtering systems, predict the utility of a specific items for specific user and based on previous user preferences or other user idea, appropriate proposals are recommended. But when some of user interest is not available, to deal with the spars problem and scalability item-based methods are used. Initially, proposed methods determine the similarity between items and then determine the items reliability to predict accuracy. In this paper, a combination of reliability and also similarities between items a new method is proposed. This new method is called trust item.

II. RELATED WORK

To how gathering information on user profiles producing two approaches have been proposed:

first approach is explicit modeling. In this method user's personal data is explicitly asked and based on his or her answers the personal profile is created (2, 3, 4, 5, 6, 9, and 10).

In the second approach, for implicit user modeling is made with minimum his or her interfere and the profile is made based on user interaction with system. These interactions are like user reviews, past queries and past emails (8, 11, and 12).

There are two approaches for how user profile is doing. First approached is crisp method. In these systems after user profile creation, users according to their profile decided to become a permanent member and after page contents are considered, pages are definitively grouped and relationships between concepts and users will also are established (3,4,11,12).

But in second approach fuzzy system is used. There is fuzzy vision to find concepts in documents, recognize user need from his or her queries and there is relationship between document and the concept. We cannot be claimed that a membership only member of a group and not interested in other groups, so these classifications is non-deterministic nature (1.7.8).

III. PROPOSED METHOD

Explicit modeling method is more exhaustive for users may or may not want to cooperate. It is also difficult to understand what user wants from his or her behavior. Despite this complexity, and to prevent and investigate questions of the user, the goal is to be learned each concept is good for user based on user behavior.

The proposed method is divided into two phases: online and offline. Online method used for model creation and offline method used for proposition and prediction. Figure one shows these steps:

978-1-4673-6490-4/13/$31.00 ©2013 IEEE

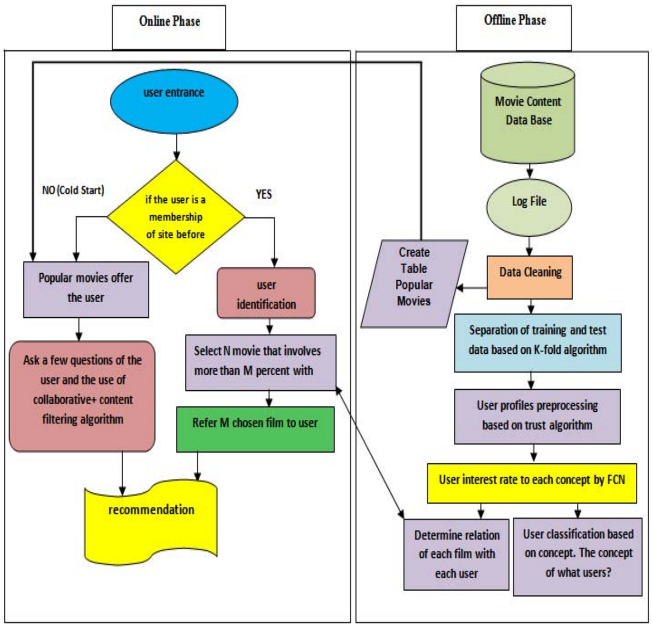

Figure 1. Proposed method in terms of offline and online

A. How to calculate trust:

Before describing the algorithm, some parameters are introduced:

Scoring Matrix User – Item

If K users U={$u_1,u_2,\ldots,u_k$} to n items I={$i_1,i_2,\ldots,i_n$} are given data explicitly, then this matrix has K*n member and it will be called A Scoring Matrix User-Item. In this matrix rows as users, columns as Items and $A_{a,j}$ as scores that user a give to item j. in these matrixes there are some entries that they haven't values, because some items don't scores with users.

User Rating Matrix Prediction – Item (P)

This matrix contains the prediction of user ratings to items. With using matrix A system could calculate previous $P_{a,i}$ for i item that scored with user a.

Matrix of user error – item

With using actual and predicted scores in the A and P matrix, we could obtain E user error matrix and with subtract user predicted score from real score E is obtained. To create E matrix, first user score to an item must be predicated. For this purpose Resnick user based prediction criterion could be changed [14]. The prediction of a user on i item, $P_{a,j}$ is obtained with this formula:

$$P_{a,i} = \overline{A_i} + \frac{\sum_{j\in N(a)} (A_{a,j} - \overline{A_i}).sim(i,j)}{\sum_{j\in N(a)} |sim\ (i,j)|} \qquad (1)$$

In this formula N(a) a set of similar items that user a scored and $A_{a,j}$ is scoring a user to j item. Also $\overline{A_i}$ and $\overline{A_j}$ are average scores of i and j items. Sim (i, j) shows the similarity between i and j items. First user predictions items on User rating prediction matrix – Item are shown. With absolute error could create user- item error matrix. Absolute error $E_{u,j}$ for pairs of actual and predicted scores <$A_{a,j}$, $P_{a,j}$> is calculated as follows:

$$e_{u,j} = |A_{u,j} - P_{u,j}| \qquad (2)$$

Reliability of an item indicates percentage of accurate predictions for an item that in each column user-item error matrix was calculated and defined as follows:

$$Confindence\ (j) = \sum_{u\in U} \frac{|E_{u,j} \cap E^r_{u,j}|}{|E_{u,j}|} \qquad (3)$$

In this formula $E_{u,j}$ is the set of predicate errors of user u on item j and $E^r_{u,j}$ are the errors that applies to requirement $e_{u,j} < \varepsilon$ and U are set of users that scored to item j. for example if a hundred errors calculate for item j and eighty of them are accurate, Confidence item j equal to 0.8.

B. predicted based on Item- Trust

CF item based approach makes similar item models that can obtain offline before proposed or anticipated offline. Given that more work is done in the offline phase, this approach could have an efficient fast online. In addition, it helps to solve the problems of distribution and scalability [15, 16]. Another advantage of the proposed method is capable of protecting against improper influences rates. In order to support fast online predictions, the amount of trust between the two items, as offline like confidence item- item matrix can be measured. Figure 2 shows the process of creating a trust matrix items – items.

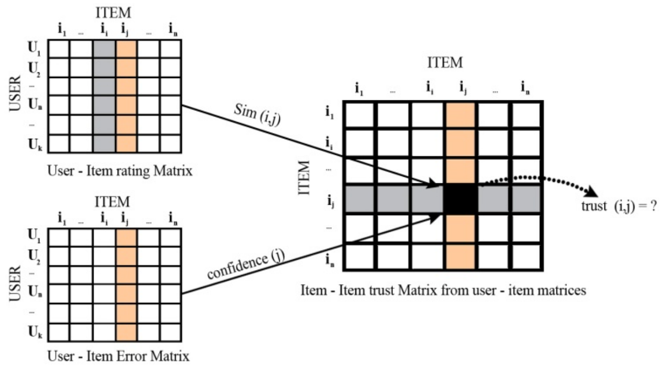

Figure 2. Create trust matrix

Trust Matrix Item – Item

Item trust model can be shows as T matrix and in this matrix rows and columns are indicated the items. Matrix element of trust items - items filled matrices by user –item matrices or A and E by using trust formula:

$$trust^{\beta}_{i\to j} = \frac{(\beta^2+1).sim(i,j).confidence(j)}{(\beta^2.sim(i,j)+confidence(j)} \qquad (4)$$

In this equation β parameter is determine to adjust weight ratio between items similarity and reliability of an item. If β=0 then $trust^{\beta}_{i\to j}$ only influenced by sim(i,j) and if β=+∞ then

$trust^{\beta}_{i\to j}$ only influenced by confidence(j). But if β=1 then both value sim (i, j) and confidence (j) are considered. The value of trust between two items is in range [0,1] and asymmetric ($trust^{\beta}_{i\to j} \neq trust^{\beta}_{j\to i}$. the appropriate value for β is chose with experimental analysis. The most important operation in CF is trying to guess the score of an item that the user is willing to give [15]. In order to compute predicted ratings of user a to item I, predicted criterion Resnick item based is used and in this computation instead of using common items sim (i, j), predicted algorithm in online phase, uses the value of the item $trust^{\beta}_{i\to j}$

$$P_{a,i} = \overline{A_i} + \frac{\sum_{j\in N(a)} (A_{a,j}\,\overline{A_i}).trust^{\beta}_{i\to j}}{\sum_{j\in N(a)} trust^{\beta}_{i\to j}} \qquad (5)$$

In this equation N (a) a K set of most similar items that user a is rated and $A_{a,j}$ is the score that user a gave to item j. also $\overline{A_i}$ and $\overline{A_j}$are the average of I and j scores.

C. Fuzzy Conceptual Networks

Users often not represented their interests accurate and reliable. Therefor fuzzy conceptual networks are a good device to represent the interest of user basis of fuzzy.

Conceptual network is a graph where each vertex represents a concept or a document and each edge is specifying the degree of association between two concepts or a concept and a document.

With this way can be discovered the concepts of an Article, although the user does not specify exactly what they are. For efficiency execution of this network conceptual matrixes are used to find the fuzzy relation rate between concepts in fuzzy conceptual networks. With computing the transitive closure of the matrix implicitly rank fuzzy relationship between concepts are obtained.

Describing document matrix

If shows the document set with P= {d_1, d_2,…, d_n} and concept set with C= {c_1, c_2,…, c_n} then the relationship between the concepts and documents has been set by an expert and shows with this matrix:

$$D = \begin{array}{c c} & \begin{array}{cccc} c_1 & c_2 & \cdots & c_n \end{array} \\ \begin{array}{c} d_1 \\ d_2 \\ . \\ . \\ . \\ d_m \end{array} & \begin{bmatrix} t_{11} & t_{12} & \square & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \end{array}$$

Figure 3. Describing document matrix

This matrix called document description matrix D. The degree of relationship between concepts and documents in document described matrix regardless of the user's interest. By multiplying document described matrix (that contains the relation between films and topics) to user described matrix (containing related topics and users) can be defined fuzzy relation between users and films with respect to user interest.

D. Processes in the proposed method

The work steps are divided into two phases, offline and online:

Offline phase includes the following steps:

1-Data cleaning.

2-Identify training and test data with K-Fold method.

3-preprocessing user profiles based on the results of the Trust algorithm, means to consider users similarity based on cosine similarity algorithm and predicts the score rate of films that don't scored by users with Trust algorithm. The output of this step is an item -user matrix that all of its elements have values.

4-Determine interest rate of each user by each concept (topic) with FCN:

$$[\,]_{user,film} * [\,]_{film,consept} = [\,]_{user,consept}$$

5-Determine any topic, including what users. In application-concept matrix, in each concept column search for users with more interest percentage rate than δ. The output of this table shows each user how interested in each concept.

6-Determine the relevance of each film to each user:

$$[\,]_{film,concept} * [\,]_{consept,user} = [\,]_{film,user}$$

The output table shows how each movie is related to each user.

Online phase includes the following steps:

1-At this stage, it is checked whether the user has already been registered or not?

2-If the user is already registered, by identifying the user, the user's interest by any means is determined with FCN.

3-Due to the relevance of each film for each user that was offline phase, n films with more λ percentage relation with user is chose.

4-Shows λ film related to user that never seen by user before.

5-But if the user is already not registered and there is no knowledge of his or her interests (Cold start problem), according to the feedback of previous users and user responses to the questions he or she is asked explicitly and the cold start problem is also solved.

IV. Evaluation

In this section, experimental results of the proposed method are provided. To evaluate performance, the proposed method compared with collaborative filtering based on item and based on user. For test data used MoveiLens that is a recommender systems research based on web. This dataset include 100000 scores that give by 943 users to 1682 films. (Item-user matrix A has 943 rows and 1682 columns). For considering proposed method accuracy item-user error matrix E is produced. Then item-item matrix T for evaluate the new method is produced. In order to measure the mean absolute error of predictions (MAE) for measure the statistical accuracy of different algorithms are used. Average absolute error for user U calculates with this formula:

$$MAUE\ (U) = \frac{\sum_{i \in I} |A_{u,i} - p_{u,i}|}{|I_u|} \qquad (6)$$

In this equation I_u is the number of items that user U scored. $\langle A_{u,i}, P_{u,i} \rangle$ real rate and predicted of user U in test data and finally average error rate for all users calculate as follows:

$$MAE = \frac{\sum_{u=1} MAUE\ (U)}{K} \qquad (7)$$

MAUE (U) average error rates for user U and K all members of the dataset.

A. required tests for parameter setting

β and similarity criterion are two important parameters in test. In evaluation tests data set with complete size and K=7 are used. K is the number of similar items with considering items.

B. to similar algorithms comparison

To evaluate item trust prediction method an item-user matrix called P for calculates confidence with close relation to similarity algorithm is used. In proposed paper (13) different similarity algorithms for implication are compared. Such as the similarity algorithm based on the correlation coefficient, algorithms based on cosine similarity, similarity algorithm based on cosine set and similarity algorithm based on the inverse correlation frequencies item. In all algorithms based on similarity criteria Risnick item is used to generate forecasts. Table one shows that prediction with cosine+iff method has better results than other quality algorithms.

TABLE I. : COMPARISON OF PREDICTION QUALITY ACHIEVED BY FIVE DIFFERENT SIMILARITY MEASURES

	Cosine	Cosine+iff	Correlation	Correlation +iff	Adjusted Cosine
MAE	0.74919	0.74248	0.75946	0.75242	0.76408

C. The sensitivity of β in reliability of items

β is a parameter used to adjust the weight and item similarity and confidence for each item is more important in trust production. Figure 4 shows change in the mean absolute error with changing β. As can be seen, prediction quality improves when B value changes from 0 to .5 and when β value tends to infinity, curve will begin to rise. So β=.5is chose as optimum value for trust formula calculation.

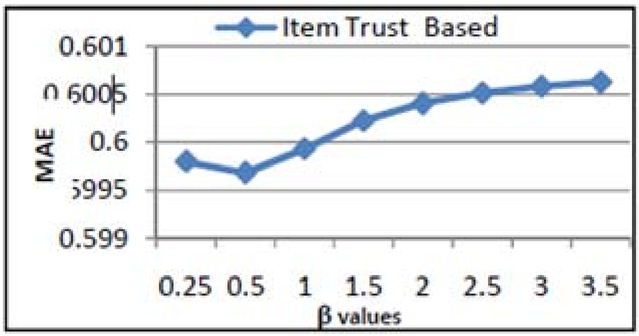

Figure 4. Sensitivity parameter b for items using trusted

D. Performance evaluation

Performance evaluation, which works on prediction quality, is based on the trust item.

E. Quality prediction based on trust item

The size of the model to predict the quality of a model-based approach is very important. If size is much larger, model is more costly. In experimental tests full-sized models are used, but the numbers of neighbors are change. Figure 5 shows laboratory results of neighbors on predicted quality. As can be seen with increasing neighbors of 2 to 7 the predicted quality is improved and then slowly decreases. The results show that proposed method, in all comparisons with different neighbors, has better accuracy than methods based on item and user.

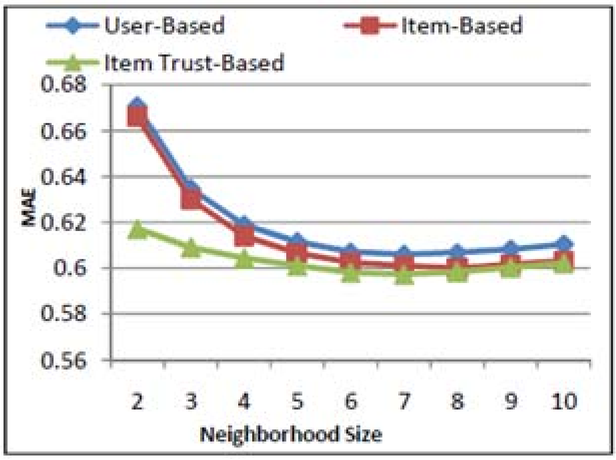

Figure 5. Comparison of methods for predicting the quality of community-based filtering and item-based and user-based trust items

F. Robustness against manipulation score rates

To assess the strength of the poor rating, 10%, 20%, 30% and 40% respectively fake rates are included in the training set and experiments were performed on full size model with number of neighbors N=7 and the results shows in table 2. In general, growth fakes score rates lead to decrease predicted quality. But trust-based collaborative filtering improves the performance in all conditions, in compare with other methods.

TABLE II. : STRENGTH OF USER-BASED COLLABORATIVE FILTERING METHODS ITEM BASED AND ITEM TRUST

unreal scors	Item Trust-based CF	Item-basedCF	User-based CF
0 %	0.5972	0.6010	0.6059
10 %	0.6437	0.6489	0.6541
20 %	0.689	0.7039	0.7148
30 %	0.7667	0.7851	0.8041
40 %	0.8034	0.8497	0.8618

So, trust-based collaborative filtering, in 4 items experiences 12% reduction on average in strength compared with the main training set(0% manipulative rates). While the average reduction strength for user-based collaborative filtering is 15% and for collaborative filtering item based is 14%. In other words, when the deceitful rates increases from 0% to 40%, the average error rate in proposed method less increased than other methods. This indicates that the proposed method in contrast with increasing rates deceitful, has less error and lead to improve strength parameters of mass filtering methods based on trust in compared to other methods. Figure 6 shows proposed method give belter offers better than other methods.

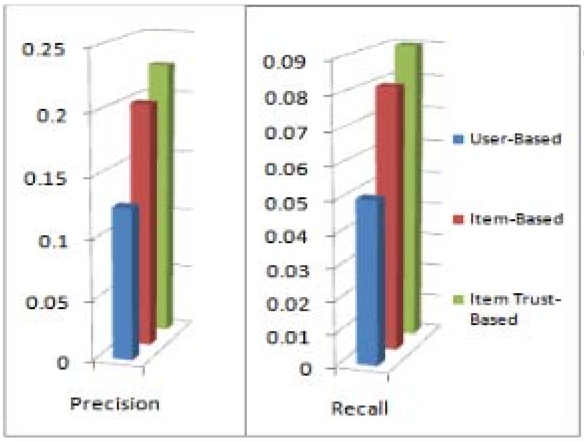

Figure 6. The accuracy of the proposed method compared with other methods in proposal movies.

V. CONCLUSION

Collaborative filtering methods on offer, is a powerful technology for users to find relevant information needs. The method proposed in this paper, a new approach to improve the quality of predictions to overcome some limitations of previous CF systems offers. In this paper, a new method to build a model, called trust item-item matrix for recommender systems based on CF is proposed. A major advantage of the proposed approach is supporting the protection of the poor score rates. Experimental Results show the proposed method in addition to improves the quality of predictions interaction with the poor dataset in compared to traditional CF algorithms is achieved.

REFERENCES

[1] S. M. Chen, Y. J. Horng and C. H. Lee, "Fuzzy Information Retrieval based onMulti-relationship Fuzzy Concept Networks", Fuzzy Sets and Systems, vol. 140, pp.183-205, 2003

[2] P. Ferragina and A. Gulli, "A Personalized Search Engine Based on Web Snippet Hierarchical Clustering", Proceedings of the World Wide Web Conference (WWW) (The Tokio, The Japan), pp. 801-810, 2005

[3] W. Alhalabi, M. Kubat and M. Tapia, "Search Engine Personalization Tool Using Linear Vector Algorithm", Proceedings of the 4th Saudi Technical Conference and Exhibition, pp. 336-344, 2006

[4] M. Radovanovic and M. Ivanovic, "CatS: A Classification-Powered Meta-Search Engine", Advances in Web Intelligence and Data Mining, Springer-Verlag, vol. 23,pp. 191−200, 2006

[5] W. Kim, L. Kerschberg, A.Scime,"Learning for automatic personalization in a semantic taxonomy based meta-search agent", Elsevier, Electronic Commerce Research and Applications 1 (2002)

[6] K .J . Kim, S .B . Cho," Personalized mining of web documents using link structures and fuzzy concept networks", Elsevier, Applied Soft Computing 7 (2007) 398−410,2005

[7] C .H .Lee,Y . H .Kim,P . K .Rhee, "web personalization expert with combining collaborative filtering and association rule mining technique", Elsevier, expert system with application 21(2001) 131-137

[8] P .Kazienko , M . Adamski,"Adrosa − adaptive personalization of web advertising", Elsevier, information sciences 177(2007) 2269-2295

[9] Y .F .kuo,L . S .Chen,"personalization technology application to internet content provider", Elsevier, Expert Systems with Applications 21 (2001) 203 – 215

[10]L. Yuen, M. Chang, Y. K. Lai and C. K. Poon, "Excalibur: a Personalized Metasearch Engine", Proc. 28th Annual International Computer Software and Applications Conference (COMPSAC'04), vol. 2, pp.49-50, 2004.

[11]J. Teevan, S. T. Dumais and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities", Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 05), ACM Press, pp.449-456, 2005.

[12]S. Souldatos, T. Dalamagas and T. Sellis, "Captain Nemo: A Meta-Search Engine with Personalized Hierarchical Search Space", INFORMATICA -LJUBLJANA- ,vol. 30, pp. 173-182, 2006

[13] Heung-Nam Kim, Ae-TtieJi, Geun-SikJo"Enhanced Prediction Algorithm for Item-Based Collaborative Filtering Recommendation",© Springer-Verlag Berlin Heidelberg 2006

[14] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In Proc. of the ACM Conf. on Computer supported Cooperative Work (1994) 175–186

[15] Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based Collaborative Filtering Recommendation Algorithms. In Proc. of the

10th Int. Conf. on World Wide Web (2001)

[16]Deshpande, M., Karypis, G.: Item-based top-N recommendation algorithms. ACM Transactions on Information Systems, Vol. 22 (2004) 143–177