A Multiresolution Approach to Recommender Systems

Gilbert Badaro
American University of Beirut
Beirut, Lebanon
ggb05@aub.edu.lb

Hazem Hajj
American University of Beirut
Beirut, Lebanon
hh63@aub.edu.lb

Ali Haddad
Yale University
Connecticut, USA
ali.haddad@yale.edu

Wassim El-Hajj
American University of Beirut
Beirut, Lebanon
we07@aub.edu.lb

Khaled Bashir Shaban
Qatar University
Doha, Qatar
khaled.shaban@qu.edu.qa

ABSTRACT

Recommender systems face performance challenges when dealing with sparse data. This paper addresses these challenges and proposes the use of Harmonic Analysis. The method provides a novel approach to the user-item matrix and extracts the interplay between users and items at multiple resolution levels. New affinity matrices are defined to measure similarities among users, among items, and across items and users. Furthermore, the similarities are assessed at multiple levels of granularity allowing individual and group level similarities. These affinity matrices thus produce multiresolution groupings of items and users, and in turn lead to higher accuracy in matching similar context for ratings, and more accurate prediction of new ratings. Evaluation results show superiority of the approach compared to state of the art solutions.

Keywords

Recommender System, Sparse Matrix, Partition Tree, Multiresolution Analysis, Haar Basis, Coupled Geometry.

1. INTRODUCTION

Recommender systems continue to capture attention due to their potentials in helping users with their daily-life decisions [1], and providing information relevant to their needs [2]. These systems are being utilized in recommending social events based on the user geographical information [3], helping users select their travel packages [4], recommending web pages [5] and solving patent maintenance related problems [6]. Moreover, in [7] the authors proposed an online system for recommending collaborators in scientific research across different domains. Traditionally recommender systems rely on ratings provided by the users on previously listed items. Existing approaches for recommender systems can be classified into: *collaborative filtering* techniques, *content-based* techniques, *hybrid models*, and *preference-based* methods [8].

NAKDD'14: The Eighth SNA-KDD wor, August 24 - 27 2014, New York, NY, USA

http://dx.doi.org/10.1145/2659480.2659501

Collaborative filtering techniques [9-12] can be divided into user-based and item-based collaborative filtering. In the user-based case, similarity between two users is computed while in the item-based case, similarity between two items is computed. Collaborative filtering can be memory based or model-based. **Content-based** techniques [8] look at the content of the items and try to retrieve features specific for a certain type of items. Textual features are usually used as features for content-based systems. **Hybrid models** were developed to overcome the limitations of collaborative filtering and content-based techniques [8-15]. Briefly, these models combine the outputs of collaborative filtering and content-based methods using for example linear combinations of predicted ratings or voting schemes. Latent factor models were also added to collaborative filtering and content-based to improve the representation of user preferences [16]. **Preference based** methods are newly introduced approaches for recommender systems that identify abstract features and relations based on the user profile which could include user's age, gender and location as implemented in [17-20].

Despite progress made in recommender systems, several challenges remain such as privacy issues [21] and evaluation metrics [22]. At the forefront, scalability of these systems remains a challenge to handle large scale number of users and items. Furthermore, existing recommender systems still face challenges in dealing with sparse data and achieving high accuracy. The issue with missing rating leads to inaccuracies when trying to match items or users for rating prediction. In this paper, we propose to address this issue of sparse user-item matrices, while achieving higher accuracies than existing systems.

2. MULTIRESOLUTION APPROACH FOR RATING PREDICTION

In this section, we explain an approach to address the sparsity problem in a user-item matrix. The approach is based on a new formulation for harmonic analysis [23] with application to user-item matrix by deriving a multiresolution transformation of the matrix similar to a wavelet transform. This transformation inherently captures the interplay between users and items at multiple levels of granularity, which enables similarity evaluation at both individual and group levels. The intuition is that similar users are likely to have similar items and similar items are likely to have similar users. While previously published hybrid recommender solutions covered aspects of the interplay between users and items, none provided the simultaneous evaluation of interplay measures at multiple resolutions, which should render rating estimation more accurate. Following the transformation, the most dominant coefficients in the new transformed space are used to reconstruct a matrix similar to the original user-item matrix, but

with estimations of previously missing ratings. Hence, the issue of sparse user-item matrix is solved. The overall approach can be represented by 8 steps as shown in Figure 1.

Step 1: In this step, the similarities among users and items are computed separately using correlation measures. The similarity measures are represented by the so called affinity matrices. For user affinity, similarity is measured between rows of the user-item matrix. For item affinity, similarity is measured between columns.

Steps 2-4: These steps constitute an iterative process that converges to two multiresolution representation of the user-item matrix. The process initially starts by deriving a multiresolution partition tree for items as described in [23] by clustering similar groups of items (columns) together at different granularity levels. The clustering is performed using diffusion distances [24]. In step 3, the interplay and similarity between users (rows) and items (columns) are measured by computing similarity between users (rows) based on the similarities across rows of the items' partition tree, i.e. rows across the multiresolution levels of the items' partition tree. This measure of interplay is called dual affinity. A sample of dual affinity computations is shown in Figure 2. Similarly, the interplay and similarity between items (columns) and users (rows) are measured by computing similarity between items (columns) based on the similarities across columns of the users' partition tree derived in step 4, i.e. columns across the multiresolution levels of the users' partition tree. This iterative approach in steps 2-4 is repeated until the partition trees converge with very little change from one iteration to the next. The results of this convergence are two multiresolution partition trees capturing the interplay: one for the users and one for the items.

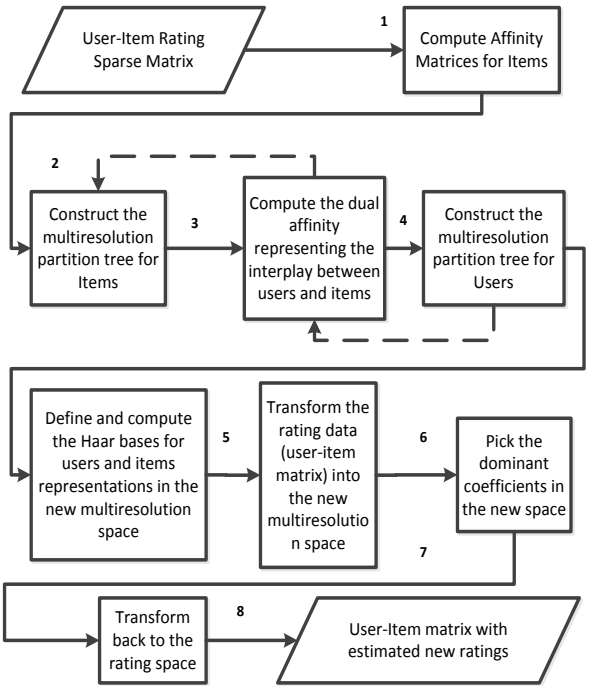

Figure 1: Overview of steps in multiresolution approach to derive new estimated ratings.

Steps 5-6: Through these two steps, the user-item matrix is transformed to the new multiresolution space with the use of the orthonormal Haar-like bases constructed from the partition trees. The product space spanned by the tensor product of the Haar-like bases can then be used to represent the original rating matrix in the new space. The orthonormal representation is constructed to represent the original user-item matrix in the new space of the multiresolution partition trees. Given a partition tree, T_Y, created from the rows of a user-item matrix M, we compute an orthonormal basis that spans the set of step functions that are constant on the nodes of the partition tree T_X, created from the columns of M, at a given level of the tree. As an example, columns of M are interpreted as step functions that are constant on the nodes at the finest level. Following the same process, a Haar-like basis can be constructed for the partition tree T_Y in relation to the columns Y.

Step 7: Similar to a Wavelet transform, this new transform can be used to efficiently compress and denoise the user-item matrix. As a result, in step 7 of the approach, the dominating coefficients in the transformed space are selected to provide the efficient representation.

Step 8: This step involves a step similar to an inverse wavelet transform. The dominant coefficients selected from the previous step are transformed back to the original user-item space. Typically, a small percentage of the coefficients are used to reconstruct the new estimated user-item matrix with the desired capture for previously missing ratings.

Since the user-item matrix is not smooth, in other words, the rating values are typically discrete, the described process (steps 1-8) of transformation and reconstruction can be iterated several times to get smoother and more accurate results. This is called a spin cycle procedure. For p iterations, the final estimated user-item matrix is calculated by averaging over the individual user-item matrices outcome from each iteration. The number of iterations can be chosen as a tradeoff between accuracy of convergence and computational complexity. For the experimentation in this paper, it was set to 5.

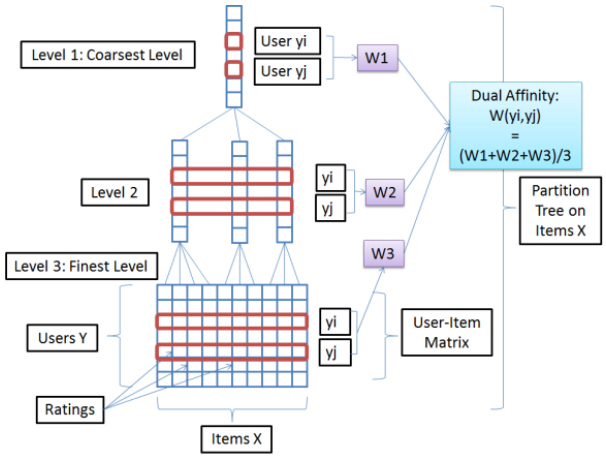

Figure 2: Example of using dual affinity from items' partition tree to compute similarity for users.

3. EVALUATION

Several experiments are conducted to evaluate the effectiveness of the proposed approach. In 3.1, we compare its accuracy against conventional user-based and item-based collaborative filtering techniques. In 3.2, we compare the accuracy against state of the art methods using larger data. In 3.3, we study the time performance of the approach and compare it to recent methods.

3.1 Comparison with Conventional Methods

The experiments described in this section are conducted on data selected from MovieLens [25], a web-based movie recommendation system that debuted in 1997. The data was collected from hundreds of users who had visited MovieLens to

rate and receive recommendations for movies. For the chosen data set, 100,000 ratings provided by 943 users to 1682 items were used for the evaluation. The data has a sparsity level of 0.94, which indicates that the matrix has 94% of its entries equal to zero which is a high degree of sparsity.

A 5-fold cross validation approach was applied (i.e. 80% training data and 20% test data). The accuracy of the proposed approach was compared to user-based collaborative filtering as stand-alone and item-based collaborative filtering as stand-alone. For fair comparisons, the algorithm was repeated 5 times, and the results were averaged consistent with the spin cycle procedure. We used mean absolute error (MAE) to measure the deviation of recommendations from their true user-specified values. For each rating-prediction pair $<p_i, q_i>$, p_i being the predicted value and q_i the correct value available in the testing data MAE is computed as follows with N being the total number of prediction pairs.

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

The proposed Harmonic Analysis approach achieved an improvement of 40% compared to user-based collaborative filtering and item-based collaborative filtering. Hence, the harmonic analysis approach has the lowest MAE compared to the two other methods.

3.2 Comparison with State of the Art Methods

This set of experiments are targeted for comparing the proposed approach with some more recent and state of the art work in the field [26], [27] and [28]. The authors in [26] present a mixed matrix factorization approach that relies on exploiting latent factors and extracting the context of the user to predict item ratings. In [27], the authors proposed an approach for improving collaborative filtering using a fuzzy clustering algorithm. For the experiments, we choose two different large datasets: the MovieLens 1M and 10M ratings.

Following the same testing process described in the previous section, the results are reported in Table 1. The new method showed, on average, accuracy improvements of around 25%, 13% and 14% compared to [26], [27] and [28] respectively.

Table 1: MAE for 1M and 10M MovieLens ratings.

Method	1M ratings	10M ratings
Multiresolution approach	**0.65**	**0.61**
Mixed Matrix Factorization [26]	0.86	0.84
Fuzzy Clustering [27]	0.72	0.71
SVD [28]	0.725	0.715

3.3 Analysis of Time Performance

To evaluate the performance of the algorithm, we evaluated the tradeoff between system accuracy and time of running the algorithm. The experiments were implemented on MATLAB installed on an Intel core i7 machine with 6GB DDR3 RAM running Windows 7 64-bit. Time measurements were collected to reflect the duration needed to run one full iteration of the algorithm and reconstruct the user-item matrix. As shown in Figure 3, the experiments indicated that the total algorithm time decreases with the number of nearest neighbors k_n used to construct the partition trees. In Figure 4, the MAE and time measurements are plotted based on varying the k_n parameter for the 10M dataset. It can be seen from this graph that the choice of k_n gives a tradeoff between accuracy and time. Also, MAE increases when k_n increases while the time required by the algorithm decreases when k_n increases. By checking the corresponding accuracies, the k_n values were chosen to be 15, 40 and 70 for the datasets 100k, 1M and 10M ratings respectively. These choices of k_n were based on the variation of time compared to the variation of MAE for each case of k_n. We chose a point where the MAE was given a higher tradeoff of accuracy versus computation time. As an example for the 10M dataset, the choice is pointed out by the arrow in Figure 4.

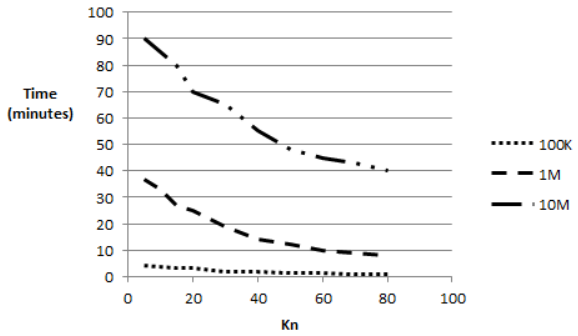

Figure 3: System Performance in terms of k_n for the three cases of Movielens dataset: 100 K, 1M, and 10M.

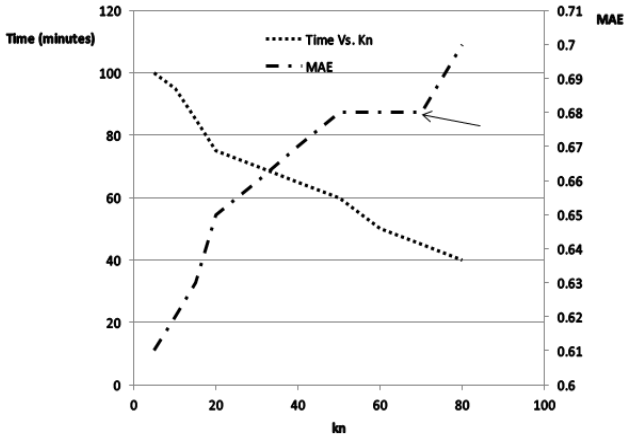

Figure 4: Performance and MAE versus choice of nearest neighbors (k_n) parameter for the 10M MovieLens dataset.

For comparison of time performance, we provide time measurements per iteration for the 1M dataset in comparison with state of the art approaches. **16**, 19, 21 and 18 are the times in minutes per iteration for our proposed approach, and the approaches in [26-28] respectively. The proposed approach provides better scalability performance.

4. CONCLUSION

In this paper, we proposed a new multiresolution approach for recommender systems based on Harmonic Analysis. New affinity measures that consider the interplay between users and items were defined for user-item matrix. The proposed approach improves the accuracy of systems with sparse user-item ratings and outperforms

conventional and state of the art methods in terms of accuracy and time performance per iteration.

5. ACKNOWLEDGMENT

This work was made possible by NPRP 6-716-1-138 grant from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

6. REFERENCES

[1] HengSong Tan; HongWu Ye, "A Collaborative Filtering Recommendation Algorithm Based on Item Classification," *Circuits, Communications and Systems, 2009. PACCS '09 Pacific-Asia Conference*, pp.694-697, May 2009.

[2] Martinez, L.; Rodriguez, R. M.; Espinilla, M., "REJA: A Georeferenced Hybrid Recommender System for Restaurants," *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences*, vol.3, pp.187-190, September 2009.

[3] Quercia, D.; Lathia, N.; Calabrese, F.; Di Lorenzo, G.; Crowcroft, J., "Recommending Social Events from Mobile Phone Location Data," *Data Mining (ICDM), 2010 IEEE 10th International Conference on* , vol., no., pp.971-976, 13-17 Dec. 2010.

[4] Qi Liu; Yong Ge; Zhongmou Li; Enhong Chen; Hui Xiong, "Personalized Travel Package Recommendation," *Data Mining (ICDM), 2011 IEEE 11th International Conference on* , vol., no., pp.407-416, 11-14 Dec. 2011.

[5] Qingyan Yang; Ju Fan; Jianyong Wang; Lizhu Zhou; , "Personalizing Web Page Recommendation via Collaborative Filtering and Topic-Aware Markov Model," *Data Mining (ICDM), 2010 IEEE 10th International Conference on* , vol., no., pp.1145-1150, 13-17 Dec. 2010.

[6] Xin Jin; Spangler, S.; Ying Chen; Keke Cai; Rui Ma; Li Zhang; Xian Wu; Jiawei Han; , "Patent Maintenance Recommendation with Patent Information Network Model," *Data Mining (ICDM), 2011 IEEE 11th International Conference on* , vol., no., pp.280-289, 11-14 Dec. 2011.

[7] Tang, J., Wu, S., Sun, J., and Su, H., "Cross-domain collaboration recommendation." In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1285-1293. ACM, 2012.

[8] Adomavicius, G.; Tuzhilin, A., "Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering 17*, pp.634-749, 2005.B.

[9] Sarwar, G. Karypis, J. Konstan, and J. Riedl. "Item-based Collaborative Filtering Recommendation Algorithms," In Proc. *of the 10th International WWW Conference*, 2001.

[10] T. Hoffman; "Latent Semantic Models for Collaborative Filtering," *ACM transactions on Information Systems*, vol.22, no.1, pp.89-115, 2004.

[11] Umyarov, A.; Tuzhilin, A.; "Improving Collaborative Filtering Recommendations Using External Data," *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on* , vol., no., pp.618-627, 15-19 Dec. 2008.

[12] Chen, Gang; Wang, Fei; Zhang, Changshui; , "Collaborative Filtering Using Orthogonal Nonnegative Matrix Tri-factorization," *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, vol., no., pp.303-308, 28-31 Oct. 2007.

[13] M. Jamali and M. Ester; ICDM 2011, Conference Tutorial, Topic: "Mining Social Networks for Recommendation," Simon Fraser University ICDM 2011, Dec. 12, 2011.

[14] Basiri, J.; Shakery, A.; Moshiri, B.; Hayat, M.Z.; "Alleviating the cold-start problem of recommender systems using a new hybrid approach," *Telecommunications (IST), 2010 5th International Symposium on* , vol., no., pp.962-967, 4-6 Dec. 2010.

[15] Claypool, M.; A. Gokhale; T. Miranda; P. Murnikov; D. Netes; M. Sartin, "Combining content-based and collaborative filters in an online newspaper," *ACM SIGIR'99, Workshop on Recommender Systems: Algorithms and Evaluation*, vol., no., Aug. 1999.

[16] Koren, Y., "Factorization meets the neighborhood: a multifaceted collaborative filtering model." *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426-434. ACM, 2008.

[17] Hornick, M.F.; Tamayo, P.;, "Extending Recommender Systems for Disjoint User/Item Sets: The Conference Recommendation Proble," *Knowledge and Data Engineering, IEEE Transactions on*, vol.24, no.8, pp.1478-1490, Aug. 2012.

[18] Konstan, J.A.; Riedl, J.; , "Recommended for you," *Spectrum, IEEE*, vol.49, no.10, pp.54-61, October 2012.

[19] Wang, C., and M. Blei, D., "Collaborative topic modeling for recommending scientific articles." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

[20] Purushotham, S., Liu, Y. and Jay Kuo, C-C., "Collaborative topic regression with social matrix factorization for recommendation systems." arXiv preprint arXiv:1206.4684 (2012).

[21] Armknecht, F.; Strufe, T., "An efficient distributed privacy-preserving recommendation system," *Ad Hoc Networking Workshop (Med-Hoc-Net), 2011 The 10th IFIP Annual Mediterranean*, pp.65-70, 12-15 June 2011.

[22] Jinoh Oh; Sun Park; Hwanjo Yu; Min Song; Seung-Taek Park; , "Novel Recommendation Based on Personal Popularity Tendency," *Data Mining (ICDM), 2011 IEEE 11th International Conference on* , vol., no., pp.507-516, 11-14 Dec. 2011.

[23] R. R. Coifman and M. Gavish, "Harmonic analysis of digital data bases," in *Wavelets and Multiscale Analysis*, Birkhäuser Boston, pp. 161-197, 2011.

[24] Coifman, R.; Lafon, S.; Lee, A.; Maggioni, M.; Nadler, B.; Warner, F. and Zucker, S., "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *In Proceedings of the National Academy of Sciences of the United States of America 102*, no. 21 pp. 7426-7431. 2005.

[25] "MovieLens Data Sets": http://www.grouplens.org/node/73, May 5, 2012 [Aug. 24 2013]

[26] Mackey, L. W.; Weiss, D. and Jordan, M., "Mixed membership matrix factorization," *In Proceedings of the 27th*

international conference on machine learning (ICML-10), pp. 711-718. 2010.

[27] Treerattanapitak, K. and Chuleerat, J., "Exponential fuzzy C-means for collaborative filtering," *Journal of Computer Science and Technology 27*, no. 3 pp. 567-576. 2012.

[28] Koren Y, Bell R. Advanced in collaborative filtering. *In Recommender Systems Handbook* (1st edition), Springer, 2011, pp. 145-186.