A hybrid recommendation system based on density-based clustering

Theodora Tsikrika, Spyridon Symeonidis, Ilias Gialampoukidis, Anna Satsiou, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute
Centre for Research and Technology Hellas
{theodora.tsikrika,spyridons,heliasgj,satsiou,stefanos,ikom}@iti.gr
http://mklab.iti.gr/

Abstract. Collaboratve filtering recommenders leverage past user-item ratings in order to predict ratings for new items. One of the most critical steps in such methods corresponds to the formation of the neighbourhood that contains the most similar users or items, so that the ratings associated with them can be employed for predicting new ratings. This work proposes to perform the combination of content-based and ratings-based evidence during the neighbourhood formation step and thus identify the most similar neighbours in a hybrid manner. To this end, DBSCAN, a density-based clustering approach, is applied for identifying the most similar users or items by considering the ratings-based and the content-based similarities, both individually and in combination. The resulting hybrid cluster-based CF recommendation scheme is then evaluated on the latest small MovieLens100k dataset and the experimental results indicate the potential of the proposed approach.

Keywords: Collaborative filtering, neighbourhood formation, hybrid recommender systems, clustering, DBSCAN, MovieLens100k

1 Introduction

Personalised recommender systems target users on the basis of their interests and preferences with the goal to suggest items (e.g., news articles, movies, music, etc.) that would appeal to them. User preferences are typically expressed through users' explicit item ratings, as well as through any other interactions that implicitly indicate users' interests, such as viewing specific items and/or commenting on them; for simplicity, we use the term "rating" to refer both to explicit ratings, as well as to implicit interactions. Such ratings, in conjunction with any additional user information (such as demographic data and explicitly provided topics of interest) constitute a user profile that, together with historical information stored in the system, is exploited by recommender systems for suggesting to users new items for which the predicted interest to them is high.

Recommendation approaches are typically divided into three broad categories: content-based, collaborative filtering, as well as hybrid approaches that combine evidence from the latter two. *Content-Based* approaches aim to recommend items that have similar content to the items the user in question (referred to as the *active user*) has rated in the past [17]. Item descriptions are typically based on textual features either automatically extracted from such items, e.g., Web pages, news articles, etc., and/or assigned by users, such as tags. Such textual representations are employed to generate user profiles that are though often noisy, due to the ambiguity inherent in natural language. Moreover, content-based systems suffer from over-specialisation, as they can only recommend items with similar content to those already rated by users in the past.

Collaborative Filtering (CF) approaches, on the other hand, provide recommendations based on the past ratings of the active user, as well as the ratings of other users in the system [5]. They are considered a particularly successful form of personalisation based on the key idea that the rating of the active user u for an item i that he/she has not encountered (and thus not rated) before would be similar to that provided by like-minded users and/or analogous to the ratings assigned by user u to similar items. As CF approaches do not take into account the actual content of the items, they can recommend items on very different topics, enabling serendipitous discoveries, provided that other users have shown interest to such diverse items. Moreover, the peer-review and word-of-mouth paradigms that form the basis of CF approaches can be considered as a more robust indicator of quality, compared to content, as they rely on the feedback of a large and potentially diverse user community, rather than only on the history of an individual user. Finally, such approaches are particularly fitting to items for which content is unavailable or difficult to obtain and/or process automatically, such as images, video, and audio.

Collaborative filtering methods are typically grouped in two broad classes: (i) neighbourhood-based and (ii) model-based methods. *Neighbourhood-based* (or *memory-based*) methods [15] leverage directly the user-item ratings stored in the system in order to predict ratings for new items. User-based neighbourhood methods estimate ratings on an item i unseen by the active user u, based on ratings for this item by other users that have similar rating patterns, referred to as neighbours, while item-based neighbourhood methods estimate such ratings based on the ratings of u for items similar to i. *Model-based* approaches, on the other hand, use these ratings to learn a predictive model which is trained using the available data, and later used for predicting user ratings to new items. While recent investigations show that state-of-the-art model-based approaches can in several cases predict ratings more accurately than neighbourhood-based methods [14], it has also been recognised that user satisfaction also depends on other factors beyond accuracy, including serendipity, justifiability, and efficiency, all of which are well served by neighbourhood-based methods [15]. Therefore, this work focuses on neighbourhood-based methods both for their simplicity and the aforementioned advantages and benefits.

One of the most central and critical steps in such methods corresponds to the formation of the neighbourhood that contains the most similar users or items, so that the ratings associated with them can be employed for predicting new ratings. Candidate neighbours are identified for each user or item using a similarity metric on the basis of the available ratings and typically the top-k neighbours with the highest similarity weights (or those with similarity weights above a certain threshold) are selected [12, 15]. Clustering approaches have also been applied for grouping the users or items, with the elements in the cluster(s) containing the user or item in question considered as its neighbours [16, 19, 1]; however, such cluster-based CF approaches have not been widely investigated.

Irrespective though of the neighbourhood formation approach that is employed, existing CF approaches typically rely on the similarities between users or items as these are estimated on the basis of the ratings information. As such they operate on a sparse space that requires a substantial number of engaged users, while they also do not allow the recommendation of items that have not been rated yet (*cold start items*). As these limitations can be ameliorated by content-based approaches, hybrid approaches that combine CF and content-based approaches have been proposed [3]. Such approaches typically combine the predictions of the two approaches to produce a single recommendation or use one approach to refine the recommendations by another.

This work proposes to perform the combination of content-based and ratings-based evidence during the neighbourhood formation step of a CF approach and thus identify the most similar neighbours in a hybrid manner. To this end, DBSCAN, a density-based clustering approach, is applied for identifying the most similar users or items by considering the ratings-based and the content-based similarities, both individually and in combination. The resulting hybrid cluster-based CF recommendation scheme is then evaluated on the latest small MovieLens100k dataset. The main contributions of this work are thus the following: (i) the proposal of hybrid neighbourhood formation on the basis of content-based and ratings-based similarities, (ii) the application of DBSCAN in a cluster-based CF scheme, and (iii) the experimental evaluation of the proposed recommender on a publicly available benchmark dataset.

This hybrid recommender scheme is proposed in the context of the activities of the PROFIT H2020 project[1] that aims to develop a platform that promotes the financial awareness and improves the financial capability of citizens and other financial market participants. In this context, the proposed recommender will be employed towards providing personalised recommendation of items, such as financial news articles, forecast reports, investment suggestions, etc.

The remainder of the paper is structured as follows. Following the discussion on related work (Section 2), the proposed approach is presented (Section 3). Then, the setting of the evaluation experiments (Section 4) and the experimental results (Section 5) are discussed, before concluding this work (Section 6).

2 Related work

Recommender systems [2] are being widely employed in multiple domains with CF methods [5] being considered a particularly successful form of personali-

[1] http://projectprofit.eu/

sation. Neighbourhood-based CF methods [15] have been extensively investigated [12] with a particular focus on the formation of an appropriate neighbourhood by considering beyond the ratings information, e.g., by selecting the most trustworthy users [20].

Clustering algorithms for neighbourhood formation in CF approaches have not been widely investigated. Some of the early approaches partitioned the user or the item space with the goal to improve the scalability using, for instance, hierarchical clustering [16] or k-means variants [19]. In a similar manner, the Normalised Cuts graph partitioning algorithm has also been used as a tool for neighbourhood selection [1]. These approaches though operate strictly on the user-item ratings. More recently, approaches that take into account additional evidence have been developed, including a multiview clustering method that iteratively clusters users on the basis of both rating patterns and social trust relationships [9], as well as a co-clustering technique that exploits item interactions and relationships, together with user community information [22].

3 Cluster-based hybrid recommender

User-based neighbourhood recommendation methods predict the rating $\hat{r}_{ui}$ by a user u for a new item i, on the basis of the ratings given to i by the user's neighbours, i.e., the users v most similar to u with respect to a similarity metric. Given that only the users who have actually rated i in the past can be considered for this estimation, the neigbourhood of user u is denoted as $N_i(u)$. The predicted rating $\hat{r}_{ui}$ can then be estimated as the average rating given to i by the neighbours v in $N_i(u)$ weighted by their similarity w_{uv} to user u and normalised. To account for user variation, i.e., to consider that users may use different rating values to quantify the same level of appreciation for an item, user ratings are typically normalised; this normalisation has shown to significantly affect performance [13].

This work applies mean-centering normalisation [12] which determines whether a rating is positive or negative compared to the mean, resulting in the following rating prediction estimation [15]:

$$\hat{r}_{ui} = \hat{r}_u + \frac{\sum_{v \in N_i(u)} w_{uv}(r_{vi} - \hat{r}_v)}{\sum_{v \in N_i(u)} |w_{uv}|} \tag{1}$$

The similarity weights w_{uv} are estimated using the Pearson Correlation Coefficient (PC) that is widely applied in CF systems and has shown robust performance despite its inherent drawbacks [10]. One of the caveats with the above is that the predicted rating can be outside of the range of the actual allowed rating (e.g., 1 to 5 stars). This is addressed by capping ratings below the lower bound and above the upper bound to the bound values.

User neighbourhoods are typically formed by considering the k users with the highest similarity weights. This work applies a clustering approach, and in particular DBSCAN, for neighbourhood formation and considers as neighbours the users that have rated i, belong to the same cluster as u, and have the k highest similarity weights. DBSCAN [6] is a density-based clustering algorithm that defines clusters as a maximal set of density-connected points based on two parameters: (i) the desired density level ϵ (= maximum radius of the neighborhood with respect to a given similarity metric, where Euclidean distance is typically used) and (ii) a lower bound $MinPts$ for the number of points in a cluster (i.e., the ϵ-neighbourhood of a point). DBSCAN is applied in this work as it can find arbitrarily shaped clusters, can deal with noise in the spatial collection of points and is thus robust to outliers, is able to extract clusters without a priori knowledge of the number of clusters, and requires just two parameters.

Based on the above, this work proposes several cluster-based CF approaches that apply different neighbourhood formation strategies; the proposed approaches are summarised in Table 1. First, DBCAN is applied on the basis of the ratings-based scores and selects as neighbours the top-k users who have rated i and belong to the same ratings-based cluster as u; Equation 1 is then applied with w_{uv} corresponding to the PC similarity weights.

Second, user clusters are formed by taking into account the content-based similarities between the items rated by the users. Each item i is represented as a feature vector x_i (using e.g., *tf.idf* weights for term-based representations of textual documents) and a user representation x_u in this feature space (referred to as *user profile*) is obtained by averaging the features of items he/she has rated I_u in the past, weighted by the actual assigned rating: $x_u = \sum_{i \in I_u} r_{ui} x_i$. DBSCAN is then applied so as to form user clusters in the content-based (CB) feature space. The selected neighbours correspond to the top-k users who have rated i and belong to the same CB-based cluster as u; Equation 1 is applied with w_{uv} corresponding to the CB similarity weights estimated using cosine similarity.

Finally, a hybrid neighbourhood formation strategy is applied by considering the intersection of users that belong to the same CB-based cluster as u and also have the highest ratings-based PC scores. In particular, the users who have rated i and belong to the same CB-based cluster as u are filtered so as to keep only the top-k who also have the highest PC scores. This allows to consider users who are similar both in terms of their rating patters, but also in terms of the content of the items they have rated. Equation 1 is then applied with w_{uv} corresponding to the PC similarity weights.

Item-based rating prediction approaches are defined in an analogous manner.

Table 1. Cluster-based CF recommendation methods

Method	Neighbourhood formation	w_{uv}
DBSCAN CF (PC)	Top-k users who have rated i & belong to the same ratings-based cluster as u	PC
DBSCAN CF (CB)	Top-k users who have rated i & belong to the same CB-based cluster as u	CB
DBSCAN CF (PC+CB)	Top-k users who have rated i & belong to the same CB-based cluster as u + have highest PC scores	PC

4 Evaluation Experiments

The proposed recommendation approaches were evaluated using a suitable dataset (Section 4.1) in a series of appropriately designed experiments (Section 4.2) with respect to well-established evaluation metrics (Section 4.3).

4.1 Dataset

The evaluation of the proposed hybrid recommendation methods requires a dataset that contains both user-item ratings and also a description of the content of these items. Given the lack of such datasets for the financial domain relevant to PROFIT, we employed one of the most established benchmark datasets, the latest publicly available small MovieLens 100k (ml-latest-small) dataset that was generated on October 17, 2016 and is publicly available at: `https://grouplens.org/datasets/movielens/`.

This dataset is provided by MovieLens, a movie recommendation service [11], and contains movies' 5-star ratings with half-star increments (i.e., 0.5 - 5.0 stars), as well as user-assigned tags to these movies, typically corresponding to single words or short phrases. It contains 100,004 ratings and 1,296 tag applications across 9,125 movies. These data were created by 671 users who were selected at random for inclusion between January 09, 1995 and October 16, 2016, while ensuring that all selected users had rated at least 20 movies; no demographic user information is included.

The dataset was split in a training set containing 80% of the data, while the rest were used for testing. The split was performed in a stratified manner on a user basis, i.e., 80% of the data associated with each user were included in the training set and the rest in the test set. This ensures that all users appear both in the training and test sets, thus avoiding the cold start user problem, and results in a test set containing 20,003 ratings for the 671 users.

4.2 Experimental Set-Up

The proposed cluster-based CF approaches can be applied as user-based or as item-based. Previous research has indicated that systems with fewer users than items may benefit more from user-based neighbourhood methods [8, 15], whereas item-based methods may achieve better performance for the opposite case [7]. Given that the selected dataset contains fewer users than items, this work evaluates the user-based version of the proposed cluster-based CF scheme.

First, rating prediction is performed according to Equation 1 where the neighbours are the k users with the highest ratings-based PC scores. Experiments are performed for $k = \{20, 30\}$ as research has shown that a number of neighbors between 20 to 50 often obtains good results [12]. These baseline approaches are compared to the proposed cluster-based CF methods listed in able 1. For the two DBSCAN parameters, $MinPts$ is set to 5 and experiments are performed for different ϵ values ranging from 10 to 100, incremented at step 10.

4.3 Evaluation metrics

The aforementioned approaches are evaluated with respect to the accuracy of ratings predictions using the well-established Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics which are defined as follows:

$$MAE = \frac{1}{|R_{test}|} \sum |\hat{r}_{ui} - r_{ui}| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{|R_{test}|} \sum (\hat{r}_{ui} - r_{ui})^2} \tag{3}$$

Both metrics compare the predicted with the actual ratings and average the error over the number of ratings that can be predicted in the test set R_{test} ; if an $\hat{r}_{ui}$ is undefined, then it is removed from consideration. Compared to MAE, RMSE assigns higher penalties to errors with larger absolute values and this has led to some criticism regarding its appropriateness for measuring model errors [21]. More recent research has shown though that a combination of metrics, including but not limited to RMSE and MAE, are often required to assess model performance [4]. Therefore, both metrics are employed in this work; since both metrics measure errors, the lower their values, the better the performance.

5 Results

The results of our experiments are presented in Table 2. As the best results for the cluster-based CF approaches were achieved for $\epsilon = 10$, we only present the MAE and RMSE values for these results.

Table 2. Experimental results for cluster-based CF approaches

		k=20		k=30	
		MAE	RMSE	MAE	RMSE
CF		0.6947	0.9113	.6927	0.9087
DBSCAN (PC)	($MinPts = 5$, $\epsilon = 10$)	0.6947	0.9113	0.6927	0.9087
DBSCAN (CB)	($MinPts = 5$, $\epsilon = 10$)	0.6919	0.9073	0.6916	0.9070
DBSCAN (PC+CB)	($MinPts = 5$, $\epsilon = 10$)	0.6947	0.9113	0.6927	0.9087

The proposed cluster-based CF approaches that employs the ratings information, i.e., DBSCAN(PC), performs equivalently to the baseline CF method. This is to be expected as neighbour selection takes place on the same space. The cluster-based CF approach that takes into account content-based information, i.e., DBSCAN(CB), manages to lower the error compared to the baseline, indicating the benefits of incorporating this form of evidence. Their combination DBSCAN (PC+CB) does not though manage to further improve the performance, indicating that the current combination does not take full advantage of both sources of evidence. An approach that would take into account users from the union of both clusters, rather than their intersection, might have been more appropriate and will be investigated as part of future work. Finally, the higher value of k achieves better results, indicating that the evidence obtained from these additional users is beneficial.

6 Conclusions

This work proposed collaborative filtering approaches that apply different neighbourhood formation strategies using the DBSCAN clustering algorithm, where the most similar neighbours are identified on the basis of clusters built using ratings-based and content-based similarities, both individually and in combination. The resulting hybrid cluster-based CF recommendation scheme was then evaluated on the latest small MovieLens100k dataset and the experimental results indicate the potential of the proposed approach and the benefits of taking into account content-based information in neighbour selection.

Acknowledgements

This work was partially supported by the European Commission by the PROFIT project (H2020-687895).

References

1. A. Bellogin and J. Parapar. Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 213–216. ACM, 2012.
2. J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.
3. R. D. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
4. T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.
5. M. D. Ekstrand, J. Riedl, and J. A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2):175–243, 2011.
6. M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996), Portland, Oregon, USA*, pages 226–231, 1996.
7. F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

8. N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. M. Sarwar, J. L. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, Orlando, Florida, USA*, pages 439–446, 1999.
9. G. Guo, J. Zhang, and N. Yorke-Smith. Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. *Knowledge-Based Systems*, 74:14–27, 2015.
10. G. Guo, J. Zhang, and N. Yorke-Smith. A novel evidence-based bayesian similarity measure for recommender systems. *ACM Transactions on the Web*, 10(2):8:1–8:30, May 2016.
11. F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2015.
12. J. L. Herlocker, J. A. Konstan, and J. Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information Retrieval*, 5(4):287–310, 2002.
13. A. E. Howe and R. D. Forbes. Re-considering neighborhood-based collaborative filtering parameters in the context of new data. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 200)8, Napa Valley, California, USA*, pages 1481–1482, 2008.
14. Y. Koren and R. Bell. Advances in collaborative filtering. In Ricci et al. [18], pages 145–186.
15. X. Ning, C. Desrosiers, and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Ricci et al. [18], pages 107–144.
16. M. O Connor and J. Herlocker. Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR workshop on recommender systems*, volume 128. UC Berkeley, 1999.
17. M. d. G. Pasquale Lops and G. Semeraro. Content-based recommender systems: State of the art and trends. In Ricci et al. [18], pages 73–106.
18. F. Ricci, L. Rokach, and B. Shapira, editors. *Recommender Systems Handbook*. Springer, 2015.
19. B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology*, volume 1, 2002.
20. A. Satsiou and L. Tassiulas. Propagating users' similarity towards improving recommender systems. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland - Volume II*, pages 221–228, 2014.
21. C. J. Willmott, K. Matsuura, and S. M. Robeson. Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752, 2009.
22. Y. Xu, Q. Yu, W. Lam, and T. Lin. Exploiting interactions of review text, hidden user communities and item groups, and time for collaborative filtering. *Knowledge and Information Systems*, pages 1–34, 2016.