Collaborative Filtering over Distributed Environment

Shlomo Berkovsky[1], Paolo Busetta[2], Yaniv Eytani[1], Tsvi Kuflik[3], Francesco Ricci[2]

[1] University of Haifa, Computer Science Department,
{slavax, ieytani}@cs.haifa.ac.il

[2] ITC-irst, Trento
{busetta,ricci}@itc.it

[3] University of Haifa, Management Information Systems Department
tsvikak@is.haifa.ac.il

Abstract. Currently, implementations of the Collaborative Filtering (CF) algorithm are mostly centralized. Hence, information about the users, for example, product ratings, is concentrated in a single location. In this work we propose a novel approach to overcome the inherent limitations of CF (sparsity of data and cold start) by exploiting multiple distributed information repositories. These may belong to a single domain or to different domains. To facilitate our approach, we used LoudVoice, a multi-agent communication infrastructure that can connect similar information repositories into a single virtual structure called "implicit organization". Repositories are partitioned between such organizations according to geographical or topical criteria. We employ CF to generate user-personalized recommendations over different data distribution policies. Experimental results demonstrate that topical distribution outperforms geographical distribution. We also show that in geographical distribution using filtering based on social characteristics of the users improves the quality of recommendations.

1 Introduction

Collaborative Filtering (CF) [1] is commonly used in many E-Commerce recommender systems to support users selecting music CDs, movies, and more [2]. CF is based on the assumption that people with similar tastes prefer the same items. In order to generate a recommendation, CF initially creates a neighborhood of users with the highest similarity to the user whose preferences are to be predicted. It then generates a prediction by calculating the normalized and weighted average of the ratings of the users in the neighborhood.

The input for the CF algorithm is a model of the user, i.e., information describing the user's preferences (interests, habits, and so on) in the form of a feature vector. This vector is matched against all other users' vectors, and *k* most similar users are selected to generate a recommendation. State of the art CF systems usually collect user models by tracking the users' past interactions with the systems, and storing this information in their local repositories. CF systems are known to suffer from two in-

herent drawbacks [3]: sparsity (lack of sufficient information about the users) and cold-start (no information about a new user or item recently added to the system).

In real life conditions, information about the users is naturally distributed among many data repositories, in a variety of domains. When integrated, these repositories could provide a recommendation, while a CF based on a single repository may fail to do so. In this work we discuss the details of operating CF over a distributed setting of data repositories and compare different distribution approaches. We facilitated the development of the above ideas using LoudVoice infrastructure. LoudVoice supports group communication in multi-agent systems, where similar service-providing agents are connected into a single virtual structure called "implicit organization" [4].

To evaluate the feasibility of our approach, we conducted several experiments. We measured the impact of different data distribution scenarios on the quality of recommendations. We compared two types of distribution representing possible real-life conditions:

- Geographical distribution - imitates a situation where information about a particular user is available only in his close vicinity. In this scenario, each LoudVoice organization represents a limited geographical area.
- Topical distribution – imitates a situation where each repository stores information related to a limited number of topics (objects types).

Experimental results show that the topical criterion is superior to the geographical criterion. Additional experiments demonstrate that applying CF using social distinction considerations (such as age, occupation and gender) improves the quality of recommendations.

The rest of the paper is organized as follows. Section *2* reviews the related works on distributed Collaborative Filtering and discusses the details of LoudVoice communication infrastructure. Section *3* discusses the possible policies of data distribution and the details of CF over the distributed environments. Section *4* presents the details of distributed CF implementation over LoudVoice. Section *5* presents experimental results. Finally, we conclude and present the directions of future research.

2 Distributed Collaborative Filtering

Collaborative filtering is probably the most familiar, most widely implemented, and most mature recommendation technique. It relies on the idea that people who agreed in the past will also agree in the future [5].

The input for the CF recommender system is a matrix of user ratings for items, where each row represents the ratings of a single user and each column represents the ratings for a single item. CF aggregates ratings of items to recognize similarities between users, and generates a new recommendation of an item by weighting the ratings of similar users for the same item [6]. The main advantage of CF is that it is completely independent of any item representation. Thus items can be recommended regardless of their contents.

2.1 Related Works

Implementing the CF algorithm in a decentralized way was initially proposed in [7]. It presents a Peer-to-Peer architecture supporting product recommendations for mobile customers represented by software agents. The communication between the deployed agents used an expensive routing mechanism based on network flooding that increased the communication overhead. An improved mechanism was proposed in [8]; however, it reduced the efficiency of the neighborhood formation phase. The work in [9] elaborated on the discussion of distributed CF. It developed a detailed taxonomy of distributed CF in recommender systems and presented different implementation frameworks for different domains of Electronic Commerce. Most of these studies did not include thorough experimentation and did not analyze the different factors that might affect the quality of the generated recommendations.

The PocketLens project [10] implemented and compared five distributed architectures for CF. It was found that no architecture is perfect, but the performance of a content-addressable mechanism [11] is close to that of a centralized CF algorithm, while the encrypted communication protocol [12] can add the essential dimension of security.

An inherent issue directly tied to the aspects of decentralized distribution is privacy. As in a decentralized setting the information resides on the client-side, individual users might restrict access to the information by deciding which other users are authorized to receive their personal information. P3P privacy policies [13], and also the work reported in [14], treat the privacy issue. It is suggested that access to the information repositories should be restricted, decreasing the likelihood of information concerning a given user being overheard by undesired parties.

Other works propose a multi-agent approach to control, and filter access to the data, depending on the user role [15], to improve privacy preservation by forming user communities. These communities acquire an encrypted aggregate user profile, representing the group as whole and not individual users [12], and employ randomized perturbation to obfuscate sensitive information about the users [16] and to minimize the possibility of acquiring such information through malicious attacks.

2.2 Self-Organized Communication Platform

In this work we employ CF over a set of distributed data repositories, where each repository acts as an independent agent. In order to minimize communication overheads, we require a platform that supports a method of communication between the relevant agents only.

LoudVoice is an efficient multi-agent communication platform based on the concept of channeled multicast [4]. Messages are sent on a channel and received by all agents that “tune” into it. Channeled multicast reduces the amount of communication needed when more than two agents are involved in a task, and allows overhearing, i.e., the ability to listen to messages addressed to others. Overhearing, in turn, enables functionality such as the collection of contextual information, pro-active assistance, and monitoring without interfering with the existing protocols.

LoudVoice has been designed to support the notion of implicit organizations. An implicit organization is a group of agents playing the same role on a given channel and willing to coordinate their actions for the sake of delivering a service. The term "implicit" highlights the fact that there is no need for a group formation phase, since joining an organization is a matter of tuning into a channel. By definition, implicit organizations are formed by agents able to play the same role. LoudVoice allows senders to address messages either to specific agents or to all agents that offer a certain service on a channel, for example providers of a particular type of information.

3 Distribution of Repositories

The neighborhood formation phase in CF finds a set of users who are similar to the user whose prediction is generated (the active user). Traditional centralized implementations typically require computing similarity between the active user and every other user in the system for the purpose of finding the set of the *K* most similar users. In a distributed environment, information about users is partitioned among different repositories. Computing similarity between users requires information stored in different and remote repositories to be combined.

In the following sub-sections we analyze two conceptually different policies for partitioning the data between the various repositories, and discuss the implications with regard to the phase neighborhood formation.

3.1 Geographical Distribution

A natural form of data distribution is "geographical distribution", where information about users is available only in their physical vicinity. For example, the reading preferences of a user are usually found in his/her local (and only local) library or bookstore. We can assume that the set of items rated by all users, in different geographical locations, is roughly similar. As each repository contains the ratings of a subset of users, geographical distribution is virtually a horizontal partitioning of the ratings matrix. Thus, the phase of neighborhood formation must comprise a search for similar users in all the repositories.

In this distribution, the set of rated items (by all users) in different repositories is fundamentally identical, and all information about a particular user is concentrated in a single repository. Therefore, to compute the similarity, the set of all the rated items of the active user should be sent to the remote repositories. Each remote repository locally computes the similarity between the active user and each of the locally stored users, and returns a "local" neighborhood. Thus, a "global" neighborhood for the active user is generated by combining all the sets of previously formed "local" neighborhoods and re-ranking the resulting set according to the users' similarity to the active user.

Comparing to a centralized CF algorithm, forming the neighborhood over geographical distribution of data repositories spreads computational load between the repositories. This occurs as each repository locally "eliminates" a portion of globally

dissimilar users, thus reducing the computational complexity of the combination of "local" neighborhoods.

In addition, geographical distribution enhances the privacy of CF, as the ratings of the users (except those of the active user) are exposed only within the boundaries of the repository (that is assumed to be more secure). Instead of all the ratings, only similarity values are transferred over the network.

3.2 Topical Distribution

A different form of data repository distribution is achieved by considering the variety of diverse domains of items (books, music, movies, and so on). This is referred to as "topical distribution" and can be considered as a vertical partitioning of the rating matrix. Each repository stores the ratings for items related to one particular domain. Thus, sets of items stored in different repositories do not completely overlap. Relying on a single domain for finding similar users might prove insufficient and might require constructing a global view of a user's ratings by combining information from the remote repositories.

In topical distribution, sets of rated items in various repositories may be different and the information about a particular user is divided among multiple repositories. Thus, there is no sense in sending the ratings of the active user items to the remote repositories, as these items might not be found there. Therefore, we base the neighborhood formation phase on the globally unique identifier (called *user-id*) of the active user (assuming that the user registers into different systems with this unique identifier only, and that the remote repository may have served the user in previous sessions).

To find the set of similar users, the active user's *user-id* is transferred over the network to the remote repositories. Each repository computes the "local" neighborhood according to ratings of its own stored items, and returns *user-ids* of potentially similar users. Similarity between the active user and the users, whose *user-ids* were returned by the remote repositories, is computed locally in the repository that initiated the recommendation process. Finally, K most similar users form the neighborhood. This type of neighborhood formation is based on the observation that the most similar users are similar in many domains and thus in a number of data repositories.

Topical distribution also enhances the privacy aspects of CF algorithm, as only local ratings are needed to compute similarity. No ratings (even those of the active user) are transferred over the network, only the *users-ids*.

3.3 Social Pre-Filtering

In addition to applying various data distribution policies, other factors could be considered for both generating a smaller neighborhood and achieving a more accurate prediction. Such considerations are based on social factors. For example, forming a "local" neighborhood in geographical distribution might limit the similarity computations to the subset of users that match the active user in one of the social criteria, such as age, occupation, social status, and so on.

Such approximation methods pose a tradeoff. On the one hand, they pre-filter potentially similar users and decrease the amount of available users for the neighborhood formation phase, thus increasing the influence of possible noise in the data. On the other hand, they might improve the accuracy of the CF, as they tend to limit the set of potentially similar users to the set of candidate users whose values for important properties are similar to those of the active user. This is actually coincides with the general notion of CF, the basis of which is a search of similar users for the purposes of building an accurate prediction.

4 Implementation details

We implemented both the topical and geographical data partitions according to the approaches presented in the previous section. We chose LoudVoice [4] to serve as an underlying communication platform. LoudVoice is an appropriate platform due to its channeled multicast capability, discussed earlier in section 2.2. This capability allowed us to handle distribution of data repositories easily, to base the distributed implementation of CF on a standardized API, and to minimize the communication overheads tied to the distribution.

Each LoudVoice channel (communication line) potentially contains a set of data repositories of either the same domain or close geographical vicinity. These data repositories are considered as an "implicit organization" of repositories. Each data repository is represented on a channel by a designated agent, whose role is to allow communication with the other agents on the channel. Organization of data sources is achieved by assigning each agent to a set of relevant LoudVoice channels, reflecting the type of the partitioning.

In addition to the agents representing data repositories, one arbitrary agent is connected to each channel and serves as a "mediator" vis a vis the other channels. This agent is connected both to his original channel and to the inter-organization communication channel. The mediators transfer requests and responses between different channels. For example, consider the structure of two LoudVoice channels "channel A" and "channel B" as illustrated in figure *1*.

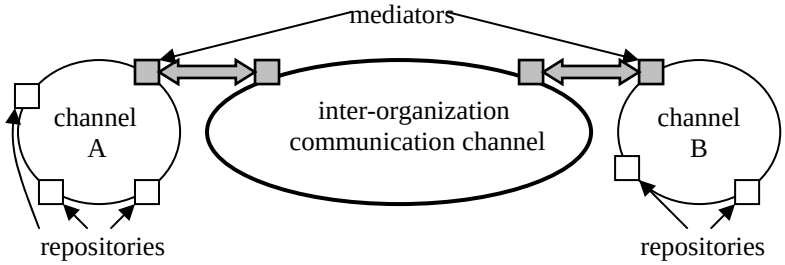

Figure 1, System Architecture over LoudVoice

5 Experimental Results

Experiments were conducted using the publicly available "1 Million Ratings" MovieLens data set [17]. It contains over one million ratings of more than *6000* different users for approximately *4000* different movies.

The first experiment aims at testing the effect of partitioning the data among multiple repositories on the quality of produced recommendations. The data was partitioned both according to geographical distribution (thus, the set of rated items in each repository is identical), and topical distribution (thus, the set of rated items might be different in different repositories).

The MovieLens dataset was partitioned among a gradually increasing number of repositories. For each number of repositories a *90%* subset of the available movie ratings was chosen to be the training set of the CF, and predictions were generated for the remaining *10%* of the ratings. The accuracy of the prediction was measured by comparing the generated prediction with the real ratings found in the data set. The metrics for the accuracy of the prediction was Mean Average Error (MAE) [18] that was computed by:

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

where N denotes the total number of the predicted items, p_i is the predicted, and r_i is the real rating on item i. To obtain statistic significance, the experiments were repeated *10* times for each number of repositories. Figure *2* illustrates the average values of MAE as a function of the number of data repositories.

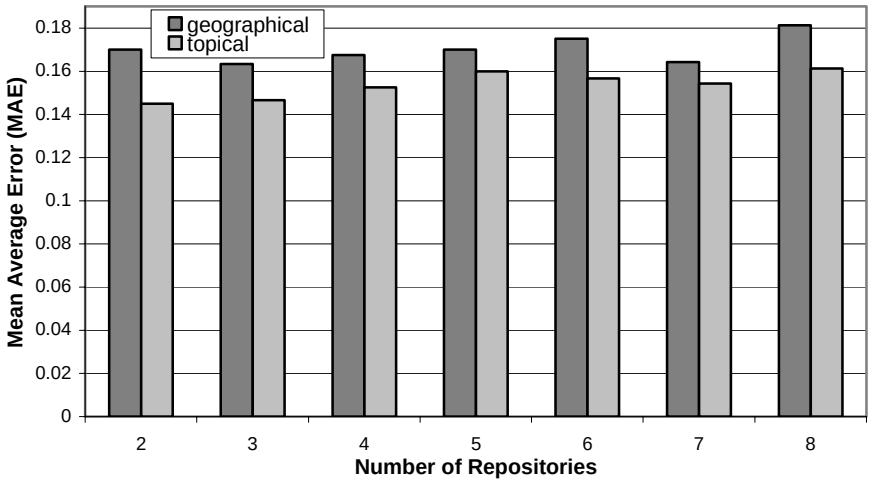

Figure 2, MAE vs. the number of repositories

The figure shows the MAE as a function of the number of repositories for both geographical (left column) and topical (right column) distributions. The MAE values are relatively low, approximately *0.14 – 0.18*, implying that generated predictions are close to the real ratings and that the MAE values are roughly indifferent to the number of repositories. The MAE values measured in the experiments are similar to those obtained in previous studies using the MovieLens dataset (initially presented in [6], and recently compared in [19]).

A comparison of the two above types of distribution shows that for any given number of repositories topical distribution slightly outperforms geographical distribution. This indicates that when the similarity is computed based only on a smaller set of relevant items, the resulting neighborhood is "closer", and as a result, the generated prediction is more accurate.

The goal of the second experiment was to measure the gains in accuracy achieved by applying social pre-filtering in addition to the geographical distribution policy. In each experiment social pre-filtering was based on one of the following social characteristics of the users: age, occupation, or gender. This information was extracted from the basic social data of the users, provided by the MovieLens dataset.

We partitioned the MovieLens dataset among a gradually increasing number of repositories. For each number, we operated each time one of the above social pre-filtering criteria: age, occupation, or gender. In this experiment also a *90%* subset of the available movie ratings was chosen to be the training set of the CF, and the predictions were generated for the remaining *10%* of the ratings. The list of potentially similar users was filtered by computing the similarity only for the users that matched the active user in the relevant social criterion. Accuracy of the prediction was computed using the MAE metrics. Figure *3* illustrates the MAE results as a function of the number of repositories.

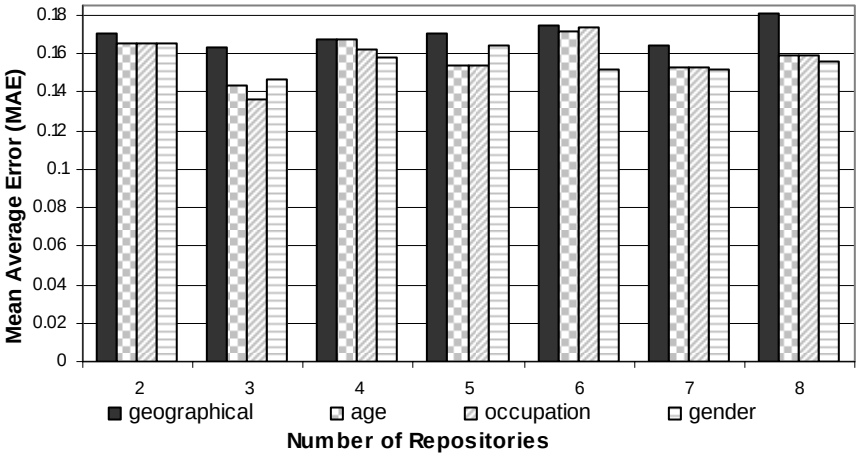

Figure 3, MAE vs. the number of repositories

Figure *3* shows that using social pre-filtering improves the MAE values, in comparison to the regular geographical distribution, although not drastically. When using the age or occupation criterion, the remaining sets of potentially similar users are relatively small. This magnifies the possible influence of noise in the data. We noticed that gender-based social pre-filtering does not act in a consistent way. Thus, experimental evidence shows that social pre-filtering generally improves prediction accuracy. However, we could not currently identify a single most contributing criterion.

6 Conclusions and Future Research

This work demonstrates the possibility of performing collaborative filtering (CF) over a distributed set of data repositories in order to resolve CF's sparsity and cold-start problems. We propose and analyze different policies for the distribution of repositories (topical and geographical partitioning). We also discuss the implementation details for each form of distribution. We suggest that preliminary filtering, based on the users' social characteristics, should be applied to improve the accuracy of the distributed CF. Though this work does not directly deal with privacy enhancement of the CF process, the proposed method of data distribution inherently contributes to solving some of the privacy issues.

The experimental results show that the accuracy of the prediction obtained from distributed CF is similar to the accuracy of state-of-the-art centralized CF systems. A comparison of two distribution policies shows that topical distribution slightly outperforms geographical distribution (regardless of the number of repositories). When the CF is preceded by social pre-filtering, the prediction accuracy increases.

A major issue that is not in the scope of this work is possible commercial competition in the E-Commerce realm. This could hamper performance by limiting cooperation and data sharing between various repositories. In the future, we plan to develop a generic model for users' cooperation and information trading.

In addition, several issues need to be addressed, such as the assumption that users' similarity remains across different organizations, and the fact that even within the same organizations terminology used by different service providers and users might differ and some kind of translation mechanism might be needed.

References

[1] J. Herlocker, J. A. Konstan, J. Riedl, "*Explaining Collaborative Filtering Recommendations*", in Proceedings of ACM Conference on Computer Supported Cooperative Work, Philadelphia, PA, 2000

[2] J. B. Schafer, J. A. Konstan, J. Riedl, "*E-Commerce Recommendation Applications*", in Journal of Data Mining and Knowledge Discovery, Vol. 5 (1/2), pp. 115-152, 2001.

[3] N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, J. Riedl, "*Combining Collaborative Filtering with Personal Agents for Better Recommendations*",

in Proceedings of the National Conference of the American Association of Artificial Intelligence, Orlando, FL, 1999.

[4] P. Busetta, A. Dona, M. Nori, "*Channeled Multicast for Group Communications*", in Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy, 2002.

[5] U. Shardanand, P. Maes, "*Social Information Filtering: Algorithms for Automating "Word of Mouth"*", in Proceedings of the International Conference on Human Factors in Computing Systems, Denver, CO, 1995.

[6] J. L. Herlocker, J. A. Konstan, A. Borchers, J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", in Proceedings of the International SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999.

[7] A. Tveit, "*Peer-to-Peer Based Recommendations for Mobile Commerce*", in Proceedings of the International Workshop on Mobile Commerce, Rome, Italy, 2001.

[8] T. Olsson, "*Decentralised Social Filtering based on Trust*", in Proceedings of the National Conference of the American Association of Artificial Intelligence Recommender Systems Workshop, Madison, WI, 1998.

[9] B. M. Sarwar, J. A. Konstan, J. Riedl, "*Distributed Recommender Systems: New Opportunities for Internet Commerce*", a chapter in "Internet Commerce and Software Agents: Cases, Technologies and Opportunities", Idea Group Publishers, 2001.

[10] B. N. Miller, J. A. Konstan, J. Riedl, "*PocketLens: Toward a Personal Recommender System*", in ACM Transactions on Information Systems, Vol. 22 (3), 2004.

[11] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker, "*A Scalable Content-Addressable Network*", in proceedings of ACM SIGCOMM Conference, San Diego, CA, 2001.

[12] J. Canny, "*Collaborating Filtering with Privacy*", in Proceedings of IEEE Conference on Security and Privacy, Oakland, CA, 2002.

[13] P3P Public Overview, http://www.w3.org/P3P/

[14] A. Kobsa, J. Schreck "*Privacy through Pseudonymity in User-Adaptive Systems*", in ACM Transactions on Internet Technology, Vol.. 3(2), pp. 149-183, 2003.

[15] L. Kagal, T. Finin, A. Joshi, "*A Policy Based Approach to Security for the Semantic Web*", in Proceedings the International Semantic Web Conference, Sanibel Island, FL, 2003.

[16] H. Polat, W. Du, "*Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques*", in Proceedings of International Conference on Data Mining, Melbourne, FL, 2003.

[17] 1 Million MovieLens Dataset, http://www.grouplens.org/data/million/

[18] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, "*Evaluating Collaborative Filtering Recommender Systems*", in ACM Transactions on Information Systems, Vol. 22(1), pp. 5-53, 2004.

[19] D. Lemire, A. Maclachlan, "*Slope One Predictors for Online Rating-Based Collaborative Filtering*", in proceedings of the SIAM Data Mining Conference, Newport Beach, CA, 2005.