Identifying User and Group Information from Collaborative Filtering Data Sets

Josephine Griffith[1], Colm O'Riordan[1] and Humphrey Sorensen[2]

[1] Dept. of Information Technology,
National University of Ireland, Galway, Ireland

[2] Dept. of Computer Science, University College Cork, Cork, Ireland

Abstract. This paper considers the information that can be captured about users and groups from a collaborative filtering data set with a view to creating user models and group models. The approach outlined defines a number of user and group features which are represented using a graph model where links exist between users and items, between users and users, and between items and items. The main focus of this paper is to extract implicit information about users and groups that exists in a collaborative filtering data set.

1 Introduction

Modern information spaces are increasingly becoming more complex where information and users are linked in numerous ways, both explicitly and implicitly, and where users are no longer anonymous, but generally have some identification and a context in which they navigate, search and browse. This offers new challenges to recommender system designers, both in capturing this information and combining it to provide for a more personalised and effective retrieval experience for a user.

The original foundations of collaborative filtering came from the idea of "automating the word of mouth process" that commonly occurs within social networks [1], i.e. people will seek recommendations on books, CDs, restaurants, etc., from people with whom they share similar preferences in these areas.

Although collaborative filtering is most frequently seen as a way to provide recommendations to a set of users, collaborative filtering data sets also allow for the analysis of social groups and of individual users within a group, thus providing a means for creating a new user model, group model or for augmenting an existing user or group model.

User modelling has had a long history in many computer science domains and traditionally user models were created based on evidence from explicit user actions. There has been a gradual change in this approach and now the focus is on building user models using implicit information gleaned from the user's interaction with a system, the user's interaction with data and information, and the user's interaction with other users.

A social network can be defined as a graph representing relationships and interactions among individuals [2]. Nodes in the graph represent individuals and the links

Griffith J., O'Riordan C. and Sorensen H. (2005).
Identifying User and Group Information from Collaborative Filtering Data Sets.
In *Proceedings of the 1st International Workshop on Web Personalisation, Recommender Systems and Intelligent User Interfaces*, pages 98-106
DOI: 10.5220/0001421600980106

between the nodes represent some relationship or relationships between individuals. Many modern social networks are found on the Internet in the form of virtual communities and the study and analysis of social networks occurs in many different fields.

A number of systems based on social networks and small world networks have been proposed for referral and recommendation [3], [4], [5], [6] and [7].

Other work linking social networks and collaborative filtering has viewed the collaborative filtering data set as a social network with the aim of analysing properties of users and items with a view to improving retrieval performance [8], [9], [10] and [11]. Aims other than solely improving retrieval performance are also explored in [8].

This paper considers the way that recommender systems bring users and groups together and considers the ways that the information from these recommender systems can be extracted to form user and group models. The motivation for this work is that although in collaborative filtering approaches, users are often clustered into groups based on finding "similar users", there is no modelling of the features of a particular group. In addition, a user generally only belongs to one group and, apart from the addition of new users, groups do not change significantly over time. This contrasts with work in the general field of user modelling where it is recognised that a user's interests will change over time and a model must be able to incorporate these changes.

The goal of the work presented in this paper is to specify the information which can be captured about users and groups given a collaborative filtering data set and to provide a model that will represent features of a user model and a group model that ultimately can be used to maintain histories of users and groups in a collaborative filtering information space. In this work the collaborative filtering data set is viewed as a graph or network. Features of users and groups are then represented in this graph. We believe that such a set of features can be used in order to develop more personalised recommender algorithms. Furthermore, the work can also be applied to the field of social networks for recommendation. In order to pursue such work it is necessary to first consider the user and group features available in collaborative filtering data sets.

Section 2 presents related work in collaborative filtering and social networks. Section 3 outlines the user and group features which can be extracted and presents the graph model. Section 4 discusses the potential usefulness of the features and approach and presents conclusions.

2 Related Work

Given a set of users, a set of items, and a set of ratings, collaborative filtering systems attempt to recommend items to users based on user ratings. Collaborative filtering systems generally make use of one type of information, that is, prior ratings that users have given to items, although some recent work has investigated the incorporation of additional information, particularly content. To date, application domains have predominantly been concerned with recommending items for sale (e.g. movies, books, CDs, restaurants) and with small amounts of text such as Usenet articles and email messages. The data sets within these domains will have different characteristics but they can be predominantly distinguished by the fact that they are both large and sparse: in a typical

domain, there are many users and many items but any user would only have ratings for a small percentage of all items in the dataset.

The problem space can be viewed as a matrix consisting of the ratings given by each user for the items in a collection, i.e. the matrix consists of a set of ratings $u_{i,j}$, corresponding to the rating given by user i to an item j. Using this matrix, the aim of collaborative filtering is to predict the ratings of a particular user, i, for one or more items previously not rated by that user.

The problem space can equivalently be viewed as a graph where nodes represent users and items. The links, or edges, between user nodes and item nodes represent a rating for the item by the user. Graph-based representations have been used for recommendation as well as being used in the social network analysis of collaborative filtering data sets.

2.1 Graph-Based Approaches for Recommendation

Several researchers have adopted graph representations in order to develop recommendation algorithms. A variety of graphs have been used: directed, two-layer, etc., and a number of graph algorithm approaches have been adopted (e.g. horting [12], spreading activation [13]).

Aggarwal et al. present *horting*, a graph-based technique where nodes represent users and directed edges between nodes correspond to the notion of predictability [12]. Predictions are produced by traversing the graph to nearby nodes and combining the ratings of the nearby users.

Huang et al. present a two-layer graph model where one layer of nodes corresponds to users and one layer of nodes corresponds to items [13]. Three types of links between nodes are represented: item-item links representing item similarity based on item information, user-user links representing user similarity based on user information, and inter-layer user-item links between items and users that represent a user's rating (implicit or explicit) of an item.

Transitive relationships between users, using a sub-set of this graph representation, is again explored in [14]. A bipartite graph is used with one set of nodes representing items and the second set of nodes representing users. The transactions of users and user feedback is modelled as links connecting the nodes between the two sets. The goal was to compare how well different collaborative filtering approaches deal with the sparsity problem and the cold start problem for new users.

2.2 Collaborative Filtering as a Social Network

As well as being used for recommendation, a collaborative filtering data set has been viewed as a social network where nodes in the network represent users and the link between users can be calculated based on the items users have rated and/or the actual ratings that users have given these items [2], [11], [9]. Rashid et al. state that "In contrast to other social networks, recommender systems capture interactions that are *formal*, *quantitative*, and *observed*" [9].

A social network can be defined as a network (or graph) of social entities (e.g. people, markets, organisations, countries), where the links (or edges) between people

represent social relationships (friendships, work collaborations, social collaborations, etc.). Recently, online relationships between people have also been used to create social networks.

A number of systems based on social networks and small world networks have been proposed for referral and recommendation. Such social networks have been built using histories of email communication [6]; co-occurrence of names on WWW pages [7]; co-use of documents by users [15]; and matching user models and profiles [3].

Rashid et al. [9] view the collaborative filtering data set as a social network where users are linked to each other. The aim is to find *influential users*. Lemire also considered influence and found that recommendation results were better if the system was not "too democratic", i.e. it was found that it was better not to penalize users with a high number of ratings [10]. In addition, Lemire considers the *stability* of a collaborative filtering system, defining stability as a property which exists if a single user in a large set does not make a difference to the results for some active user.

Mirza et al. also induce a social network from a collaborative filtering data set where connections between users are based on co-ratings of the same items [8]. They define a *hammock jump* as a connection between two users in the network that will exist if the users have at least w items co-rated (where w is defined as the hammock width). Herlocker et al. refer to this measure as a *significance weighting* whereby they devalue the correlation value between two users if this correlation value has been calculated based on only a small number of co-rated items [16].

Palau et al. analyse collaborative data sets using a number of properties of social networks [11]. These include: size, density, degree centrality, network centrality, clique membership and factions.

3 User and Group Features

This section specifies the implicit information about users (Section 3.1) and groups (Section 3.2) that can be extracted from a collaborative filtering data set. A graph model is presented in Section 3.3 which can be used to represent the explicit and implicit user and group information.

3.1 User Models

A user model is defined which consists of a number of features, with the values of all features in the range $[0, 1]$ and all calculated in comparison to all other users in the data set.

For a user a the following is considered:

– *rated* is the number of items rated by user a in comparison to the number of items rated by all users:

$$\frac{i_a - i_{min}}{i_{max}}$$

where i_a is the number of items rated by user a; i_{min} is the minimum number of ratings a user gives and i_{max} is the maximum number of ratings a user gives. The higher the value the more ratings a user has given.

- *avg-rating* and *st. dev* are the average score given to items by user a and the standard deviation of user a in comparison to the average score given to items by all users (and associated standard deviations) respectively.
- *liked* is the number of items liked by user a (counting the items whose value is greater than or equal to the average rating of user a as *liked* [17]) calculated by:

$$\frac{liked_a - liked_{min}}{liked_{max}}$$

  where $liked_a$ is the number of items liked by user a; $liked_{min}$ is the minimum number of items liked by a user and $liked_{max}$ is the maximum number of items liked by a user.
- *disliked* is the number of items disliked by user a (counting the items whose value is less than the average rating of user a as *disliked* [17]) calculated by:

$$\frac{disliked_a - disliked_{min}}{disliked_{max}}$$

  where $disliked_a$ is the number of items disliked by user a; $disliked_{min}$ is the minimum number of items disliked by a user and $disliked_{max}$ is the maximum number of items disliked by a user.
- *avg-item-popularity* is the popularity of the items rated by a in comparison to the popularity of these items across all user ratings for those items.
- *N-common-raters* is the number of users who have rated N items in common, choosing some constant value for N.
- *influence* is a measure of how influential a user is in comparison to other users (see below for description).

Apart from *influence*, the above user model features have simple calculations. As also considered in [9] and in [8], *influence* is defined here by using measures from social network theory, in particular, *degree centrality* and *distance centrality*. In order to calculate these, the data set is viewed as a social network where nodes represent users and the values of weights on edges between users are based on user ratings. Generally a social network graph may only consider the presence or absence of a link between nodes or may consider the strength of the link (or relationship) among nodes. Positive correlation values based on user ratings can also be used as the edge weights and can be used to indicate the strength of the relationship between two user nodes [9].

Degree centrality is measured by counting the number of links a node has to other nodes. A high degree centrality value indicates the level of connection with other users and a node can be considered *central* if it has a higher degree than any of the other nodes whereas a node with a low degree is isolated from most of the other users in the network [18].

Distance centrality is measured by calculating the shortest paths between nodes. Distance centrality gives an indication of the "power" of a node in terms of its distance to all other nodes. For example, in a star network one node is maximally close to all other nodes while all the other nodes are maximally distant from each other.

Another centrality measure of possible interest is *betweenness centrality* which would give an indication of "information control", i.e. the number of times a node is

between two other nodes when these two nodes have no direct link between them. This may be useful in analysing if the network is dominated by a few very central nodes as these nodes essentially become "hubs" and thus without these nodes the network may be disconnected. A star network is an example of a network containing one node with a very high *betweenness centrality* value.

3.2 Group Models

A number of different group models can be extracted from the collaborative filtering data set. The general aim of finding group models is to find the portions of the network where users are more closely linked to each other than to other users.

From a recommendation point of view, generally a user group for an active user a is defined as the user nodes which are two steps, or links, away from the user node a, possibly where links above a certain threshold weight are only considered. Some work has also considered nodes which are more than two links away from the active node as neighbour nodes (e.g. when considering transitive relationships [12], [14], [13]).

We consider groups from the perspective of an individual user a: when a is the active user the group model tells us about the users which influence a; when a is member of some active user's group the group model tells us about the influence a exerts on other users in a group.

When a is the active user, we are interested in:

1. The common items that the group has rated.
2. The number of nodes in the user's group (where the link weights must be greater than some threshold and we only consider nodes X links away from active node a).
3. The clustering coefficient of the group, i.e. a measure of how connected the neighbours of a are to each other. For example, considering the case where X is 2 so that only user nodes 2 links away from a are considered, the resulting network is a star network with a as the hub if no links between the neighbours are considered. This network has a clustering coefficient of 0. If a network has a clustering coefficient of 1 it means that all of a's neighbours are connected to each other. The clustering coefficient can be calculated by:

$$\frac{actual}{possible}$$

where $actual$ is the number of actual links between neighbour nodes and $possible$ is the number of possible links which can exist between neighbour nodes. If the links are undirected then the number of possible links that can exist between n nodes is:

$$possible = \frac{(n^2 - n)}{2}$$

Thus the formula becomes:

$$\frac{actual \times 2}{(n^2 - n)}$$

When user a is a member of some active user's group we are interested in:

1. The number of groups to which a belongs.
2. The "closeness" of a to the active user, i.e. the weight between a and the active user.
3. The clustering coefficient of a, i.e. the number of other users that a is connected to in the group (note that a will have to be connected to at least one node, the active user node).

In this section we have outlined a number of features which can be extracted from the collaborative filtering data set. These features can form part of a user and group model.

3.3 Representing Properties of Users and Groups using a Graph Model

In our representation the collaborative filtering problem space is viewed as a graph with a set of user nodes, a set of item nodes and sets of weighted edges which connect nodes.

Initially user nodes and item nodes are connected via a weighted edge, the value of which is a function of the rating given to an item by a user to indicate the level of like or dislike for that item.

Apart from some scaling of the rating value users have given to items, this graph is a direct mapping of the data in the matrix representation to a graph representation of the data. Therefore the information on user models and group models is not represented explicitly. To represent this information explicitly, additional edges are added to the graph. These additional edges represent further relationships between users and items, relationships between items and relationships between users. For example, a relationship exists between commonly rated items; between highly rated items; between users who have rated a set of the same items; between users with high standard deviations; between users with low standard deviations; etc. The value of the weights on the edges connecting these nodes is a function of the value obtained from the features listed in the previous sections.

Depending on the user of interest, or the group of interest, only some of the edges may be considered and only portions of the graph may be traversed. A spreading activation search approach [13] is being used to highlight items, users and groups where the features identified and presented in this paper are used to determine and constrain the spreading activation.

The graph model can be used for recommendation with or without the additional information from the features identified. In addition, the graph model can be used to help inform improved, more personalised recommendations and provide useful information and feedback to users. The graph model can also be used to explain more fully the reasons for good and bad recommendations.

4 Conclusions and Future Work

In this paper we have reviewed work in collaborative filtering, social networks and graph-based recommendation, highlighting the similarities between the work. Traditional collaborative filtering approaches do not adopt a graph-based representation although some graph-based approaches have recently been developed and some researchers

in the social network domain have adopted collaborative filtering approaches. In this paper, we have defined features of users and groups that can be identified from a collaborative filtering data set and that can be of use in providing more personalised recommendations. These features are represented using a graph model. We believe that these features are of use in defining and formalising algorithms for recommendation in collaborative filtering and social networking domains.

Ongoing work involves experimental evaluation of the usefulness of the graph model presented and the identified user and group features. This will involve developing and testing graph-based recommendation algorithms for collaborative filtering and comparing these with more traditional collaborative filtering approaches. As briefly mentioned, a spreading activation search approach is being used to highlight items, users and groups for recommendation to users.

In addition, future work involves demonstrating that the graphs built from collaborative filtering data sets are structurally similar to small world networks. This will strengthen the case for the application of graph-based recommendation algorithms to modern social network communities.

References

1. Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating word of mouth. In: Proceedings of the Annual ACM SIGCHI on Human Factors in Computing Systems (CHI '95). (1995) 210–217
2. Barnes, J.: Social Networks. MA: Addison-Wesley (1972)
3. Vivacqua, A., Lieberman, H.: Agents to assist in finding help. In: ACM Conference on Computers and Human Interface (CHI-2000). (2000)
4. McDonald, D., Ackerman, M.: Expertise recommender: a flexible recommendation system and architecture. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work. (2000) 231 – 240
5. Krulwich, B., Burkey, C.: The contactfinder: Answering bulletin board questions with referrals. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence. (1996)
6. Schwartz, M., Wood, D.: Discovering shared interests using graph analysis. Communications of the ACM **36** (1993) 78 – 89
7. Kautz, H., Selman, B., Shah, M.: Referral web: combining social networks and collaborative filtering. Communications of the ACM **40** (1997) 63 – 65
8. Mirza, B., Keller, B., Ramakrishnan, N.: Studying recommendation algorithms by graph analysis. Journal of Intelligent Information Systems **20** (March 2003) 131 – 160
9. Rashid, A., Karypis, G., Riedl, J.: Influence in ratings-based recommender systems: An algorithm-independent approach. In: SIAM International Conference on Data Mining. (2005)
10. Lemire, D.: Scale and translation invariant collaborative filtering systems. Information Retrieval **8** (2005) 129–150
11. Palau, J., Montaner, M., Lpez, B.: Collaboration analysis in recommender systems using social networks. In: Cooperative Information Agents VIII: 8th International Workshop, CIA 2004. (2004) 137 – 151
12. Aggarwal, C., Wolf, J., Wu, K.L., Yu, P.: Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In: Proceedings of the Fifth ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'99), San Diego, CA. (1999) 201–212

13. Huang, Z., Chung, W., Chen, H.: A graph model for e-commerce recommender systems. Journal of the American Society for Information Science and Technology **55** (2004) 259–274
14. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems **22** (2004) 116–142
15. Mukhopadhyay, U., Stephens, L., Huhns, M., Bonnell, R.: An intelligent system for document retrieval in distributed office environments. Journal of the American Society for Information Science **37** (1986)
16. Herlocker, J., Konstan, J., Riedl, J.: An empirical analysis of design choices in neighbourhood-based collaborative filtering algorithms. Information Retrieval **5** (2002) 287–310
17. Lemire, D., Maclachlan, A.: Slope one predictors for online rating-based collaborative filtering. In: Proceedings of SIAM Data Mining (SDM'05). (2005)
18. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks **1** (1979) 215–239