Data Mining Techniques for Personalization

Charu C. Aggarwal, Philip S. Yu
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Abstract

This paper discusses an overview of data mining techniques for personalization. It discusses some of the standard techniques which are used in order to adapt and increase the ability of the system to tailor itself to specific user behavior. We discuss several such techniques such as collaborative filtering, content based methods, and content based collaborative filtering methods. We examine the specific applicability of these techniques to various scenarios and the broad advantages of each in specific situations.

1 Introduction

In recent years, electronic commerce applications have gained considerable importance because of the importance of the web in electronic transactions. The Web is rapidly becoming a major source of revenue for electronic merchants such as Amazon, and Ebay, all of which use some form of personalization in order to target specific customers based on their pattern of buying behavior.

The idea in personalization in web business applications is to collect information about the customer in some direct or indirect form and use this information in order to develop personalization tools. Thus, an electronic commerce site may continue to keep information about the behavior of the customer in some explicit or implicit form. The two typical methods for information collection at an E-commerce site are as follows:

- **Explicit Collection:** In this case, the customer may specify certain interests, which reflect his desire of buying certain products. Example of such methods include collaborative filtering methods in which customers specify ratings for particular products and these ratings are used in order to make peer recommendations.

- **Implicit Collection:** In implicit collection, the E-commerce merchant may collect the buying behavior of the customer. In addition, the browsing or other bahavior may also be captured using the web or trace log. There are several limitations to the nature of implict collection because of privacy requirements of users.

This paper is organized as follows. In the next section, we will discuss collboarative filtering which is the technique of using the explicit information in order to make peer recommendations. In subsequent sections, we will discuss more indirect ways of using data mining techniques on the customer behavior in order to capture his patterns. Finally, we will discuss an overview of the general applicabilities of these techniques to different scenaries.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Collaborative Filtering Techniques

In collaborative filtering techniques, the idea is appropriate for e-commerce merchants offering one or more groups of relatively homogeneous items such as compact disks, videos, books, software and the like. Collaborative filtering refers to the notions of multiple users "sharing" recommendations in the form of "ratings", for various items. The key idea is that collaborating users incur the cost (in time and effort) of rating various subsets of items, and in turn receive the *benefit* of sharing in the collective group knowledge. For example, they can view predicted ratings of other items that they identify, see ordered lists of items whose predicted ratings are the highest and so on. The explicit ratings entered by the user are typically done so in some kind of user interface which we will refer to as a profiler dialog.

The importance of collaborative filtering arises from the fact that the number of possible items at an E-commerce site is far greater than the number of items that a user can be expected to rate in a reasonable amount of time. In fact, the catch in collaborative filtering is to give users enough motivation so that a sufficient number of items are rated. Typically, it is known that given the choice, only $1-2\%$ of the users are likely to rate items, and even among them only a small number of items can be rated. In order to implement a site in which every user rates items, E-commerce merchants typically provide incentives such as free email [**?**] or other products, which may then be used to elicit greater user behavior. Once a greater level of user behavior has been achieved, the ratings of various items may be predicted for a given user by creating peer groups. In this case, the idea is to construct groups of users having a similar behavior and using their *aggregate* behavior in order to predict the ratings for a given user-product combination which has not been explicitly specified in a profiler dialog [**?**, **?**]. Recent techniques have also been developed[**?**] which build peer groups in an even more general way by using groups of people which have either positively or negatively correlated behavior.

Some examples of useful questions which could be resolved by a recommender system are as follows:

- Show user i the projected rating for item j, assuming that the item has not already been rated by that user and assuming that the algorithm is able to meaningfully estimate the item's rating. If the user has already rated the item, that item is shown instead.

- Show user i an ordered list of up to M (as yet unrated) items from a subset $\mathcal{I}$ which are projected to be liked the most, subject to the constraint that each of them has a projected rating of at least some minimum value r. This subset may be chosen implicitly or explicitly in a number of ways. For example, it may consist of a set of promotional items.

- Show user i an ordered list of up to M (as yet unrated) items from a subset $\mathcal{I}$ which are projected to be liked the least subject to the constraint that each of them has a projected rating of at most some maximum value r.

- Select an ordered list of up to M users from a subset $\mathcal{J}$ each of which has a projected rating of at least r that are projected to like item j the most. This is a query which is useful to the E-commerce merchant in order to generate mailings for specific sets of promotion items.

Even though collaborative filtering is a very effective technique, its shortcomings are clear; it is difficult to collect a sufficient amount of user feedback to begin with. For such cases, it may be more desirable to pick techniques which deal with data corresponding to actual user behavior. Such methods are referred to as content based techniques.

3 Content Based Systems

In content based systems, the feedback of other users is not relevant while deciding the recommendations for a given user. The idea is to characterize the user behavior in terms of his *content* of accesses; a new attribute

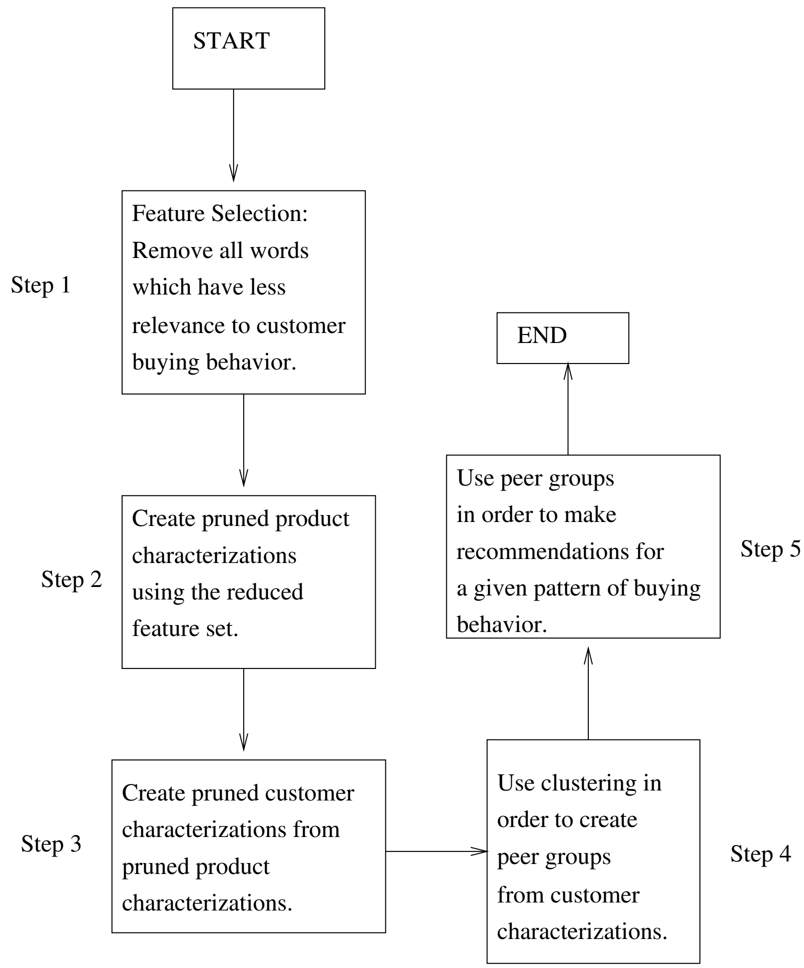

Figure 1: Content Based Collaborative Filtering

system corresponding to user behavior. For example, while tracking the user behavior at an E-commerce site, the content of his behavior may refer to the set of text documents which have been accessed by him.

At the same time, the set of products at an E-commerce site may be categorized or clustered based on the content of the products. For a given user, one may then categorize his behavior by using IR similarity techniques in order to compare the content of the cluster and the content of user access patterns [**?**]. The results of this comparison may be used in order to make product recommendations. The idea here is to create a content taxonomy and characterize each user by the content category of the products that he purchased. Similar techniques may be used in order to make web page recommendations by analyzing the web pages which were browsed.

Content based techniques have the advantages of directness, simplicity, and redundancy of user feedback, but they lack the sophistication of collaborative filtering systems which analyze the behavior of entire peer groups for making recommendations. For this purpose, it is useful to examine and analyze the behavior of content based collaboratuve systems which do not require user feedback; yet use the concept of user peer groups for making recommendations.

4 Content Based Collaborative Filtering Systems

In this section, we will discuss an application of clustering in order to build content based collaborative filtering systems. Content based collaborative filtering systems are relevant in providing personalized recommendations at an E-commerce site. In such systems, a past history of customer behavior is available, which may be used for making future recommendations for individual customers. We also assume that a "content characterization" of products is available in order to perform recommendations. These characterizations may be (but are not restricted to) the text description of the products which are available at the web site. The key here is that the characterizations should be such that they contain attributes (or textual words) which are highly correlated with buying behavior. In this sense, using carefully defined content attributes which are specific to the domain knowledge in question can be very useful for making recommendations. For example, in an engine which recommends CDs, the nature of the characterizations could be the singer name, music category, composer etc., since all of these attributes are likely to be highly correlated with buying behavior. On the other hand, if the only information available is the raw textual description of the products, then it may be desirable to use some kind of feature selection process in order to decide which words are most relevant to the process of making recommendations.

We will now proceed to describe the overall process and method for performing content-based collaborative filtering. The collaborative filtering process consists of the following sequence of steps, all of which are discussed in Figure **??**.

(1) **Feature Selection:** It is possible that the initial characterization of the products is quite noisy, and not all of the textual descriptions are directly related to buying behavior. For example, *stop words* (commonly occuring words in the language) in the description are unlikely to have much connection with the buying pattern in the products. In order to perform the feature selection, we perform the following process: we first create a customer characterization by concatenating the text descriptions for each product bought by the customer. Let the set of words in the lexicon describing the products be indexed by $i \in \{1, \ldots, k\}$, and let the set of customers j for which buying behavior is available be indexed by $j \in \{1, \ldots, n\}$. The frequency of presence of word i in customer characterization j is denoted by $F(i, j)$. The fractional presence of a word i for customer j is denoted by $P(i, j)$ and is defined as follows:

$$P(i, j) = \frac{F(i, j)}{\sum_{j \in \text{All customers}} F(i, j)} \tag{1}$$

Note that when the word $i = i_0$ is noisy in its distribution across the different products, then the values of $P(i_0, j)$ are likely to be similar for different values of j. The *gini index* for the word i is denoted by $G(i)$, and is defined as follows:

$$G(i) = 1 - \sqrt{\sum_j P(i, j)^2} \tag{2}$$

When the word i is noisy in its distribution across the different customers, then the value of $G(i)$ is high. Thus, in order to pick the content which is most discriminating in behavioral patterns, we pick the words with the lowest gini index. The process of finding the words with the lowest gini index is indicated in Step 1 of Figure **??**.

(2) **Creating Customer Characterizations:** In the second stage of the procedure, we create the customer characterizations from the text descriptions by concatenating the content characterizations of the products bought by the individual consumers. To do so, we first prune the content characterizations of each product by removing those features or words which have high gini index. Then we create customer characterizations by concatenating together these pruned product characterizations for a given customer (Step 3).

(3) **Clustering:** In the subsequent stage, we use the selected features in order to perform the clustering of the customers into peer groups. This clustering can either be done using unsupervised methods, or by supervision from a pre-existing set of classes of products such that the classification is directly related to buying behavior.

(4) **Making Recommendations:** In the final stage, we make recommendations for the different sets of customers. In order to make the recommendations for a given customer, we find the closest sets of clusters for the content characterization of that customer. Finding the content characterization for a given customer may sometimes be a little tricky in that a weighted concatenation of the content characterizations of the individual products bought by that customer may be needed. The weighting may be done in different ways by giving greater weightage to the more recent set of products bought by the customer. The set of entities in this closest set of clusters forms the *peer group*. The buying behavior of this peer group is used in order to make recommendations. Specifically, the most frequently bought products in this peer group may be used as the recommendations. Several variations of the nature of queries are possible, and are discussed subsequently.

We have implemented some of these approaches in a content-based collaborative filtering engine for making recommendations, and it seems to provide significantly more effective results than a content based filtering engine which uses only the identity attributes of the products in order to do the clustering.

Several kinds of queries may be resolved using such a system by using minor variations of the method discussed for making recommendations:

(1) For a given set of products browsed/bought, find the best recommendation list.

(2) For a given customer and a set of products browsed/bought by him in the current session, find the best set of products for that customer.

(3) For a given customer, find the best set of products for that customer.

(4) For the queries (1), (2), and (3) above, find the recommendation list out of a prespecified promotion list.

(5) Find the closest peers for a given customer.

(6) Find the profile of the customers who will like a product the most.

Most of the above queries (with the exception of (6)) can be solved by using a different content characterization for the customer, and using this content characterization in order to find the peer group for the customer. For the case of query (6), we first find the peer group for the content characterization of the current product, and then find the dominant profile characteristics of this group of customers. In order to do so, the quantitative association rule method [**?**] may be used.

5 Summary

In this paper, we discussed several implicit and explicit techniques for performing personalization at an electronic commerce site. The explicit methods for personalization include user-feedback and ratings, whereas the implicit behavior include the observation of user buying behavior. Each of these techniques have their own drawbacks:

- There are limitations to how much effort a user is willing to put in in order to explicitly provide his preferences using methods such as a profiler dialog. Typical response rates for schemes in which ratings may be provided on a voluntary basis are only $1-2\%$. Furthermore, often such information may be too sparse to be of too much use.

- There are limitations to how much one can track user behavior without violating privacy concerns that many people may have. Often, for E-commerce sits which work without the use of a registration process have no way of tracking behavior in an effective and acceptable way.

Given the limits on tracking the online user behavior, it is often easier to collect user behavior at a batch level, since the data collection process is not hindered by the online nature of the user actions. For such cases many data mining techniques such as associations, clustering, and categorization [**?**, **?**, **?**, **?**] are known.

References

[1] Aggarwal C. C. et. al. "Fast Algorithms for Projected Clustering." *Proceedings of the ACM SIGMOD Conference of Management of Data*, 1999.

[2] Aggarwal C. C., Yu P. S. "Finding Generalized Projected Clusters in High Dimensional Spaces." *Proceedings of the ACM SIGMOD Conference on Management of Data*, 2000.

[3] Aggarwal C. C., Wolf J. L., Wu K.-L., Yu P. S. "Horting Hatches an Egg: Fast Algorithms for Collborative Filtering." *Proceedings of the ACM SIGKDD Conference*, 1999.

[4] Aggarwal C. C., Gates S. C., Yu P. S. "On the merits of building categorization systems by supervised clustering." *Proceedings of the ACM SIGKDD Conference*, 1999.

[5] Agrawal R., Imielinski T., and Swami A. "Mining association rules between sets of items in very large databases." *Proceedings of the ACM SIGMOD Conference on Management of data,* pages 207-216, 1993.

[6] Chen M. S., Han J., Yu P. S. Data Mining: An overview from the database perspective. *IEEE Transcations on Knowledge and Data Engineering*.

[7] www.amazon.com

[8] www.ebay.com

[9] Freund Y., Iyer B., Shapiro R., Singer Y. "An efficient boosting algorithm for combining preferences."*International Conference on Machine Learning,* Madison, WI, 1998.

[10] Greening D., "Building Consumer Trust with Accurate Product Recommendations.", Likeminds White Paper LMWSWP-210-6966, 1997.

[11] Resnick P., Varian H. "Recommender Systems", *Communications of the ACM*, Volume 40, No. 3, pages 56-58, 1997.

[12] Shardanand U., Maes P. "Social Information Filtering: Algorithms for Automating Word of Mouth." *Proceedings of the CHI'95,* Denver CO, pp. 210-217, 1995.

[13] Srikant R., and Agrawal R. Mining quantitative association rules in large relational tables. *Proceedings of the 1996 ACM SIGMOD Conference on Management of Data.* Montreal, Canada, June 1996.

[14] www.hotmail.com