Department of Information Technology

technical report NUIG-IT-160900

Collaborative Filtering

J. Griffith (NUI, Galway)
C. O' Riordan (NUI, Galway)

Collaborative Filtering

Josephine Griffith and Colm O' Riordan
IT Centre, NUI, Galway

September 16, 2000

Abstract

This paper presents an overview of the field of collaborative filtering, a set of techniques which attempt to assist with the problem of information overload by selecting information based on recommendations (both implicit and explicit) made by other people.

This technique differs from most previous attempts in that the content of the items is not used as a factor in the filtering process.

This paper covers the motivations for adopting a collaborative approach, the main techniques and previously developed systems.

1 Introduction

Techniques to deal with information overload are becoming increasingly prevalent. The vast majority of such techniques and related systems attempt to overcome the problems of information overload by automating the analysis of the content of online documents. Documents are retrieved for a user based on a similarity measure between the representations of user's information need and documents.

Recently, newer techniques have been developed to improve upon and augment content-based approaches. These collaborative filtering systems attempt to approximate word-of-mouth recommendations. Items are not retrieved on content-based analysis but rather on recommendations by other people.

This paper reviews this relatively new field of research. In Section 2, content and collaborative filtering techniques are compared and contrasted. Section 3 details approaches to collaborative filtering, outlining the most commonly used algorithms and the related metrics used to ascertain the usefulness of the algorithms. A summary of available results is presented in Section 4, while Section 5 addresses some pertinent issues, namely difficulties that arise in obtaining ratings from users and dealing with potential sparseness of data and information. Section 6 describes common applications of collaborative filtering including details of some well-known collaborative-filtering systems.

2 Background

The ever-increasing popularity of the Internet has led to an influx of users and, consequently, a huge increase in the volume of available on-line data. Such data is available through web sites, ftp sites, mailing lists and Usenet newsgroups. This increase has led to a situation where users are swamped with information and have difficulty sifting through the reams of material, much of which is not relevant to them. This scenario is commonly referred to as the problem of *information overload*.

The vast majority of the information is in an unstructured or semi-structured format so traditional database techniques are not sufficient in aiding users.

This insufficiency has led to the development of information retrieval (IR) and information filtering (IF) as important fields.

The fields of IR and IF have attempted to provide solutions to the problem of information overload. These systems typically view the task as selecting from a set of documents, a subset of these documents, which are relevant to a user interest.

The difference between IR and IF is usually taken to be a difference in the nature of the data and the nature of the user request. With IR, the data set is relatively static and the user request is a one-off query. With IF, the data set is dynamic and the user request is a profile representing a long term interest. For an overview of the differences and similarities between retrieval and retrieval systems the user is directed to [4].

Newer systems and theories are being developed every year with results presented in conferences such as TREC[9].

According to Malone[13] there are three classes of filtering techniques: cognitive, based on the content of articles (which has received the most attention in the past), social (or collaborative), based on human judgments (which is the focus of this paper) and economic, based on the cost of producing and reading items.

2.1 Content-Based Filtering

Developed systems range from naive approaches involving simple string matching augmented with Boolean and proximity operators to more advance techniques using the vector space model, latent semantic indexing (LSI), probabilistic models (inference and belief networks) and connectionist networks.

The vector space model [15] represents documents as weighted vectors and uses a cosine formula to evaluate similarity. Turtle and Croft have used inference networks to implement comparisons[17]. Newer models borrowing ideas from the AI field include Kwok's connectionist network[11] and Belew's AIR system[2][3]. Techniques such as LSI attempt to recognise latent interconnections occurring between terms[7].

The effectiveness of the content-based approach is constrained by inherent difficulties of natural languages.

2.2 Collaborative Filtering

Collaborative Filtering is not based on analysis of the content of the document set but on the premise that "people with similar interests in the past will have the same interests and preferences in the future"[12]. Collaborative filtering systems and recommender systems attempt to exploit this information to predict users' interests.

Given a set of users, a set of items, and a set of ratings, systems attempt to recommend items to users based on prior ratings. The collaborative filtering systems essentially automates the "word of mouth" process.

2.3 Comparison of Content Filtering to Collaborative Filtering

Collaborative filtering offers a number of advantages over content-based filtering. The most obvious difference between the two types of filtering is that content filtering fails to capitalise on the knowledge and opinions of people who have previously accessed documents in the document set. With collaborative filtering, attributes such as quality, clarity, presentation style, and not just content, can be taken into account. These factors are not considered in content-based filtering.

In addition, content filtering can only be applied to textual document sets whereas collaborative filtering can be applied to both textual and non-textual items, e.g. images, sound, movies, programs, etc.

The problems of polysemy (words with the same spellng and different meanings depending on the context) and synonmyny (words that are spelled differently but have the same meaning) which often degrade the performance of content-based filtering systems, can be alleviated with collaborative filtering. Also, an item may be recommended in which the user has not explicitly expressed an interest but which nonetheless the user may find interesting (serendipity).

3 Approaches to Collaborative Filtering

The problem space can be viewed as a matrix consisting of the ratings of each user for the items in the document set, i.e., the matrix consists of a set of ratings $u_{i,j}$, corresponding to the rating by user i for 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 in the document set.

The steps involved in the prediction of these ratings for a given user i are:

- Select a set of users with similar interests/preferences to user i, i.e., users who have similar ratings for items as user i.
- Predict recommendations for user i from the set selected in step 1, i.e., if these users rated an item j highly, this item will be recommended to user i.

The various techniques which can be used to perform these two steps are discussed in the following sections.

3.1 Algorithms

Neighbourhood-based algorithms are the most commonly used approach in collaborative filtering. A subset of users is chosen based on their similarity with a current or active user. Such methods comprise three main steps:

1. Users with similar tastes to the active user are selected - calculate user correlation.
2. A subset of these users is selected as a set of predictors - neighbourhood selection.
3. A prediction is computed from the ratings of these selected neighbours - generate a prediction.

Step 1: Calculate user correlation:
Various techniques can be used to calculate the user correlation. These include:

1. *Pearson correlation:* a weighted average of deviations from the neighbours' mean is calculated. This approach was used in the original GroupLens system[14].
2. *Constrained Pearson correlation:* A variation on Pearson correlation where the deviation from the median of available rating values, rather than the deviation from the mean, is chosen in calculating correlations. This was used in the Ringo system[16].
3. *The Spearman rank correlation:* similar to Pearson correlation but uses ranking as opposed to explicit rating values, thus giving greater independence of the range of ratings.
4. *The Vector similarity:* uses the cosine measure between the user vectors to calculate correlation.
5. *Entropy-based uncertainty measure:* which uses conditional probability techniques[16].
6. *Mean-square difference algorithm:* the mean square difference between each pair of users is calculated; the smaller the difference the higher the correlation.

Step 2: Neighbourhood Selection

Given the correlation values, techniques are now needed to determine the number of neighbours

to select. In order to calculate a given set of neighbours the following techniques can be used:

1. *Correlation thresholding*—where all neighbours with absolute correlations greater than a specified threshold are selected. Selecting a high threshold means that only good correlates will be selected thereby giving more accurate predictions. It may happen that very few neighbours will have such a high correlation and as a result it may not be possible to generate meaningful predictions for some items.
2. *Best-n correlations*—where the best n correlates are picked. Picking a large value of n may result in too much noise for those with high correlates whereas picking a small n can cause poor predictions for users with low correlates.

Step 3: Generate a prediction

Once the neighbourhood has been generated the predictions can be produced. Techniques which can be used include:

1. Compute the weighted average of user ratings using the correlations as the weights. This weighted average makes an assumption that all users rate items with approximately the same distribution. This approach was used in Ringo[16].
2. The weighted mean of all neighbours' ratings is computed. Rather than take the explicit numeric value of a rating, a rating's strength is interpreted as it's distance from a neighbour's mean rating. This approach attempts to account for lack of uniformity in ratings. This approach was used in GroupLens[14].
3. To account for the differences in the range of users' ratings, a 'z-score' can be calculated for each rating which assigns significance to a rating as a function to the rating value, its distance from the neighbour's mean and the degree of variance in the neighbour's ratings.

3.2 LSI and SOM

Billsus and Pazzani[6], describes an alternative technique which combines a learning algorithm coupled with singular value decomposition (SVD). SVD (an approach also used in the LSI content filtering system) is used to reduce the dimensionality of the matrix. The initial matrix can be decomposed into 3 matices: $A = U\sigma V^T$ where U and V are composed of orthonormal vectors that define left and right singular values of A. σ is a diagonal matrix. The highest k singular values are maintained together with the corresponding rows and columns in U and V^T. From these three reduced matrices, A' an approximation of the original matrix A can be derived. This approximation of the original exploits latent interconnections between data items. A detailed description is available in [5] and [7].

The vectors in this reduced matrix are then used to train a neural network (a feed-forward network with sigmoidal activation levels), which is then used to generate predictions.

3.3 Metrics

The main metrics used to test the usefulness of a collaborative filtering algorithm are:

- *Coverage:* a measure of the ability of the system to provide a recommendation on a given item.
- *Accuracy:* a measure of the correctness of the recommendations generated by the system.

Coverage is usually computed as a percentage of items for which the system was able to provide a recommendation.

Accuracy is usually calculated by comparing the ratings generated by the system to user-provided ratings. The accuracy is usually presented as the mean

absolute error between ratings and predictions[16].

An alternative approach is to use ROC sensitivity, a measure of the ability of the system in a decision support environment. Typically, the value of the rating is not that important—we are more interested in whether it is a good or a bad rating. Usually ROC sensitivity is measured by plotting *sensitivity* against $(1 - specificity)$. *Sensitivity* is the probability of a good item being returned by the system as such, *specificity* is the probability of a *poor* item being acurately identified.

4 Results

Most experiments to test collaborative filtering algorithms adopt a similar approach. A known collection of ratings by users over a range of items is decomposed into two disjoint subsets. The first set (usually the larger) is used to generate recommendations for items corresponding to those in the smaller set. These recommendations are then compared to the actual ratings in the second subset. The accuracy and coverage of a system can be thus be ascertained.

Empirical analysis exists for quite a few of the algorithms and techniques heretofore mentioned. The majority deal with which correlation algorithm provides the best results. Results[10] show that Spearman and Pearson perform with roughly the same accuracy and that these two techniques outperform vector based comparison and mean square difference approaches.

It has also been shown that using the mean rating for an item is less effective than using the average deviation from the mean rating. No results indicate that using z-scores to account for degrees of variance result in any improvement.

5 Other Issues

While a range of techniques and systems have been successfully developed to provide accurate recommendations in a range of domains a number of issues still remain to be resolved.

One such issue is that of obtaining a sufficient volume of ratings to avoid a high degree of sparseness which will ultimately lead to low accuracy and poor coverage.

The majority of systems require users to explicitly rate items. If a large number of items exist it is likely that the task could prove too cumbersome and time-consuming for a user, resulting in an incomplete set of ratings thereby adversely affecting the performance of the system.

An alternative is to attempt to gather implicit ratings by modelling users' behaviour. This approach may prove useful in providing a sufficient number of ratings. It also removes the user burden of having to supply many ratings. The obvious pitfall associated with implicit ratings is that assumptions in the user model may prove false resulting in a decrease in accuracy.

Difficulties also arise with respect to determining the size of the neighbouhood. Too large a neighbourhood results in increased coverage but decreased accuracy. Conversely, too small a neighbourhood results in increased accuracy at the price of decreased coverage.

6 Applications and Systems

There has been a marked increase in the number of systems utilising collaborative filtering techniques. These have mainly fallen into two categories:

- Information filtering and retrieval

In these systems the items being recommended by the system are documents (emails, postings to Usenet News etc.). Users provide recommendations regarding which documents are relevant. Information retrieval systems that utilise collaborative filtering include Tapestry[1] and Grouplens[8].

Tapestry

The concept of collaborative filtering originated with the Tapestry project at Xerox PARC[8]. It allows users to explicitly annotate the electronic documents that they read. These annotations can be a textual comment or a Boolean rating. Users can then retrieve documents based on other users' opinions of the document as well as the content of the documents.

Tapestry suffers from a number of flaws including:

- the formulation of the collaborative relationships remains the task of the user.
- requires explicit user interaction.
- lack of privacy, as users know who made recommendations.

Grouplens

Grouplens is collaborative filtering system applied to Usenet news. With this system users rate articles on a numerical scale of 1 to 5. Correlations between different user ratings are calculated. Explicit user interaction is required. Experiments showed that a large number of users and their ratings were required to achieve effective performance.

- Electronic Commerce
  In these systems collaborative filtering is used to personalise the on-line shopping experience. The advantages for both the consumer and producer are apparent—time saved for the consumer, more focused selling for the producer thereby increasing probability of sales. Many online stores are beginning to adopt these techniques. For example:

  amazon: which allow people to share ratings and reviews on different books.

  dejanews: which allow recommendations on computer equipment.

  levis: who tried to recommend items to users based on past purchases.

7 Conclusion

Collaborative filtering techniques recommend items to users based on the ratings of items received from users with similar tastes to the current user. This differs from traditional approaches to Information Retrieval (IR) and Information Filtering (IF) where items are retrieved from a document set based on content alone.

This paper provides an overview of the field of collaborative filtering by discussing the need for IR and IF techniques and the difference between content-based and collaborative-based filtering. Collaborative filtering approaches and algorithms are then discussed together with the metrics that are used to test and compare these algorithms.

A summary of experimental results using the various algorithms is presented. Additional issues which must be considered with collaborative systems are outlined. Finally, a synopsis of the main systems using collaborative filtering techniques is presented.

References

[1] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[2] R. K. Belew. *Adaptive Information Retrieval: Machine Learning in Associative Networks.* Phd thesis, Univ. Michigan, CS Department, 1986.

[3] R. K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn

about documents. *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 1989.

[4] N. Belkin and B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(2), December 1992.

[5] M. Berry, S. Dumais, and G. O. Brien. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573,595, 1995.

[6] D. Billsus and M. Pazzani. Learning collaborative information filters. *AAAI Workshop on Recommender Systems*, pages 24–28, july 1998.

[7] S. Deerwester, S.Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[8] D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–69, 1992.

[9] D. Harman. Overview of the fourth text retrieval conference (trec 4). *TREC-4 Proceedings*, 1995.

[10] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. *SIGIR*, pages 230–237, 1999.

[11] K. L. Kwok. A neural network for probabilistic information retrieval. *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 1989.

[12] P. Maes. The agent network architecture (ana). *SIGART Bulletin*, 2(4):115–120, 1991.

[13] Malone, Grant, Turbak, Brobst, and Cohen. Intelligent data sharing systems. *Communications of the ACM*, 30(5):390–402, 1987.

[14] P. Resnick, N.Iacovou, M. Suchak, P. Bergstrom, and J. Reidl. Grouplens : An open architecture for collaborative filtering of netnews. *Proceedings of ACM 1994 Conference on CSCW*, pages 175 – 186, 1994.

[15] G. Salton. *Automatic Text Processing: The transformation, Analysis, and Retreival of Information by Computer*. Addison-Wesley, 1989.

[16] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating word of mouth. *Proceedings of the Annual ACM SIGCHI on Human Factors in Computing Systems (CHI '95)*, pages 210–217, 1995.

[17] H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. on Info. Systems*, 3, 1991.