Unified Filtering by Combining Collaborative Filtering and Content-Based Filtering via Mixture Model and Exponential Model

Luo Si
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lsi@cs.cmu.edu

Rong Jin
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
rongjin@cse.msu.edu

ABSTRACT
Collaborative filtering and content-based filtering are two types of information filtering techniques. Combining these two techniques can improve the recommendation effectiveness. The main problem with previous research is that the content information and the rating information are not combined in an integrated way. This paper presents a unified probabilistic framework that allows the mutual interaction between these two types of information. Experiments have shown that the new unified filtering algorithm outperforms a pure collaborative filtering approach, a pure content-based filtering approach and another unified filtering algorithm.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval]: Filtering**

General Terms
Algorithms

Keywords
Unified filtering

1. INTRODUTION
Collaborative Filtering (CF) [1,3] and Content-Based Filtering (CBF) [2] are two techniques that help users find out the most valuable information. Content-based filtering analyzes the item content representations to make recommendations, while collaborative filtering utilizes rating information to calculate the similarities between test user and training users for rating prediction. Clearly, content-based filtering and collaborative filtering compensate each other by using different types of information. Unified filtering (UF) technique takes advantage of both the content information and the rating information. This strategy is helpful to improve the recommendation accuracy.

There have been a couple of studies on combining content based filtering and collaborative filtering [2,4]. Many of them used the linear combination strategy to combine the prediction scores of CF and CBF. The deficiency with these approaches is that each type of information is treated separately and their correlation is not fully explored.

CIKM '04, November 8--13, 2004, Washington, DC, USA.

This paper presents a unified probabilistic framework to better utilize these two types of correlated information. We use a mixture model for the unified filtering approach, whose goal is to cluster users and items into groups of similar behavior. Both the user rating information and the item content information are utilized to build accurate clusters. Specifically, content information helps us build more accurate clusters for items, and more accurate item clusters results in better clustering of users, which provides more accurate similarity measurement between the item contents. Following this idea, we extend a model-based collaborative filtering approach by combining mixture model and exponential model to incorporate item content information.

2. PRIOR RESEARCH
For content-based filtering, there are many valuable methods such as logistic regression that is used in this work. For collaborative filtering, model-based methods group similar users and items together to make the prediction. Flexible mixture model (FMM) [3] was proposed and has been shown to be superior to several other collaborative filtering approaches.

Unified filtering approach combines the content information and rating information together to improve the prediction accuracy. Most previous work applies CBF and CF independently to generate prediction and the results are combined linearly. Yu et al. [4] extended the linear combination approach through a Bayesian framework called collaborative ensemble learning (CEL). Each user is associated with a probabilistic text categorization model. Then, the user content models are combined to predict ratings for test users. The main deficiency with those approaches is that they do not fully explore the correlation between content information and rating information. This is dangerous if a single CBF or CF model has poor performance and may result in poor combination output.

3. NEW UNIFIED FILTERING METHOD
Compared with the FMM mixture model for collaborative filtering, we use the content information to build better item clusters and user clusters. Our strategy is to train a discriminative model of predicting item class Z_x by the content information of item x as $P(Z_x \mid \vec{d}_x)$ instead of a generative model to avoid the bias problem of various item content lengths. Formally, the joint probability of a rating tuple that user y assign item x of rating r is defined as:

$$P(x,y,r) = \sum_{Z_x, Z_y} P(x)P(y)P_\theta(Z_x \mid x)P(Z_y \mid y)P(r \mid Z_x, Z_y) \qquad (1)$$

Where Z_x and Z_y represent the item class and the user class respectively. The item class model $P_\theta(Z_x \mid \vec{d}_x)$ is calculated by the following exponential model as:

$$P_\theta(Z_x \mid x) = \frac{\exp(\theta_{z_x,1} w_{x,1} + \theta_{z_x,2} w_{x,2} \cdots + \theta_{z_x,|v|} w_{x,|v|})}{\sum_{z_x'} \exp(\theta_{z_x',1} w_{x,1} + \theta_{z_x',2} w_{x,2} \cdots + \theta_{z_x',|v|} w_{x,|v|})} \quad (2)$$

Where $w_{x,i}$ is the normalized term frequency for the ith word feature of item x and $\vec{\theta}_{z_x}$ are the parameters that determine the importance of words to different item classes.

Annealed expectation and maximization method is used for training. In E step, the posterior probabilities are estimated as:

$$P(z_x, z_y \mid x_{(l)}, y_{(l)}, r_{(l)}) = \frac{(P(z_x \mid x_{(l)}) P(z_y \mid y_{(l)}) P(r_{(l)} \mid z_x, z_y))^b}{\sum_{Z_x, Z_y} (P(Z_x \mid x_{(l)}) P(Z_y \mid y_{(l)}) P(r_{(l)} \mid Z_x, Z_y))^b} \quad (3)$$

Where b is the temperature variable. Then the model is updated with respect to the posterior probabilities. For example, the user class probabilities are updated as:

$$P(z_y \mid y) = \frac{\sum_{l: y_{(l)} = y} \sum_{z_x} P(z_x, z_y \mid x_{(l)}, y_{(l)}, r_{(l)})}{P(y) * N\_T} \quad (4)$$

Where N_T is the total number of training ratings. The item class probabilities $P(Z_x \mid \vec{d}_x)$ are first estimated from rating information in a similar way, which is further refined by maximizing the discriminative model with respect to the content model as:

$$L_{\tilde{P}}(\theta) = \sum_{z_x, x} P(z_x \mid x) P(x) \log P_\theta(z_x \mid x) \quad (5)$$

Therefore, we can compute appropriate weight parameters that best fit the distribution $P(z_x \mid x)$ from the viewpoint of content information. In return, those weights will be used to better adjust item classes and user classes in the next iteration.

To make recommendation for a test user, a plug-in method is used to rerun the above training procedure with all model parameters fixed except for the new user class probability. Furthermore, we can use a similar strategy as FMM [3] to calculate the posterior rating probabilities and finally the prediction can be made.

4. EXPERIMENTS

Each-Movie[1] was used as testbed in this work, in which user ratings range from zero to five. 1234 movies were selected because their contents (text descriptions) can be extracted from the Internet Movie Database[2]. 1000 most common words were selected as the word features. 1000 users with more than 20 ratings were chosen, which reduces total number of ratings to 61357. The evaluation metric used in our experiments was the mean absolute error (MAE), which is the average absolute deviation of the predicted ratings from the actual ratings.

We compared our new unified filtering algorithm of combining mixture model and exponential model (UFME) with the pure content-based filtering approach (CBF) by logistic regression text categorization, pure collaborative filtering approach (CF) by FMM and another unified filtering algorithm as collaborative ensemble learning (CEL). Different configurations of different number of training users and different amount of given items from test users are explored. The results are shown in Table 1.

Table 1. MAE Results for four filtering algorithms on EachMovie. Four algorithms are pure content-based filtering (CBF), pure collaborative filtering (CF), unified filtering by collaborative ensemble learning (CEL) and unified filtering by combining mixture model and exponential model (UFME).

Training Users Size	Num of Items Given	CBF	CF	CEL	UFME
50	0	1.43*	1.21	1.24	**1.19**
	5	1.23	1.14	1.16	**1.11**
	10	1.21	1.13	1.14	**1.10**
	20	1.20	1.12	1.12	**1.09**
100	0	1.43*	**1.17**	1.22	**1.17**
	5	1.23	**1.08**	1.13	**1.08**
	10	1.21	1.07	1.11	**1.06**
	20	1.19	**1.05**	1.09	**1.05**

It can be seen from the results that the UFME unified filtering approach outperforms both the CBF and the CF approaches. Its advantage over the CF approach is more significant in the case of small amount of rating information (50 training users). Furthermore, the UFME approach is better than the CEL unified filtering approach, which we attribute to its integrated modeling and training of the content and rating information.

5. CONCLUSION

This work proposes a unified filtering framework that combines the content and rating information in an integrated manner. Specifically, the content information is incorporated into a mixture model through a conditional exponential model for the item class distribution. Empirical studies have shown that the new unified filtering algorithm outperforms the pure collaborative filtering, the pure content-based filtering methods and an alternative unified filtering approach.

6. REFERENCES

[1]. J. S. Breese, D. Heckerman & C. Kadie. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence.*

[2]. P. Melville, R. Mooney & R. Nagarajan. (2002). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of Conference on Artificial Intelligence*.

[3]. L. Si & R. Jin. (2003). Flexible Mixture Model for Collaborative Filtering. In *Proceedings of the 20th International Conference on Machine Learning.*

[4]. K. Yu, A. Schwaighofer, V. Tresp, W. Y. Ma & H. J. Zhang (2003). Collaborative ensemble learning: combining collaborative and content-based information filtering via Hierarchical Bayes. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence.*

[1] http://research.compaq.com/SRC/eachmovie

[2] http://www.imdb.com/