LDA-BASED PERSONALIZED DOCUMENT RECOMMENDATION

Te-Min Chang, Department of Information Management, College of Management, National Sun Yat-sen University, Kaohsiung, Taiwan, R.O.C., temin@mail.nsysu.edu.tw

Wen-Feng Hsiao, Department of Information Management, National Pingtung Institute of Commerce, Pingtung, Taiwan, R.O.C., wfhsiao@mail.npic.edu.tw

Abstract

Accompanying with the rapid growth of Internet, people around the world can easily distribute, browse, and share as much information as possible through the Internet. The enormous amount of information, however, causes the information overload problem that is beyond users' limited information processing ability. Therefore, recommender systems arise to help users to look for useful information when they cannot describe the requirements precisely.

The filtering techniques in recommender systems can be categorized into content-based filtering (CBF) and collaborative filtering (CF). Although CF is shown to be superior over CBF in literature, personalized document recommendation relies more on CBF simply because of its text content in nature. Nevertheless, document recommendation task provides a good chance to integrate both techniques into a hybrid one with the aim to enhance the overall recommendation performance.

The objective of this research is thus to propose a hybrid filtering approach for personalized document recommendation. Particularly, latent Dirichlet allocation to uncover latent semantic structure in documents is incorporated to help users obtain robust document similarity in CF. Two experiments are conducted accordingly. The results show that our proposed approach outperforms other counterparts on the recommendation performance, which justifies the feasibility of our proposed approach in applications.

Keywords: Recommender systems, Content-based filtering, Collaborative filtering, Hidden topic analysis, Latent Dirichlet allocation

1 INTRODUCTION

Due to the explosive growth of information over the Internet, there is more and more information disseminating and distributing through this new channel. The large amount of information, however, results in a big challenge for users who desire to find relevant information from the huge repository. This is commonly referred to as the information overload problem due to human's limited information processing ability. Consequently, recommender systems arise to assist users in retrieving useful information efficiently in such situations that they cannot describe their requirements precisely. Recommender systems suggest users the desired information through the analyses of their past preferences or the preferences of like-minded people to the users.

Filtering techniques in recommender systems can be categorized into content-based filtering (CBF) and collaborative filtering (CF). Content-based filtering techniques compare the new information with an active user's[1] profile of past interest to predict whether he/she is interested in the new information, whereas collaborative filtering techniques look for like-minded people of the active user and recommend what is interested among those people to him/her.

Most of the information circulated in the Internet is in the format of text. The issue of personalized document recommendation therefore arises and becomes essential. Due to the text content nature, the primitive approach to document recommendations is by means of CBF method. However, a possible direction is to incorporate content-based features into CF similarity computation to obtain the additive effect of both CBF and CF while minimizing their possible individual drawbacks.

Recently, several latent topic discovery techniques such as probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) appear that serve as dimension reduction approaches to exploring text content features by revealing latent topics of each document from the document repository. With latent topics uncovered, we can derive document similarity more precisely to comprehend users' requirements and make more relevant recommendation to them.

The objective of this research is thus to propose a hybrid filtering approach for the task of personalized document recommendation. Particularly, latent Dirichlet allocation is employed in our study to uncover the semantic structure hidden in the documents that users have read, which includes the word distributions over the latent topics and the latent topic distributions over documents. We can then utilize LDA results in such ways as measuring document similarity based on latent topic distributions. It is desired that the proposed hybrid approach can compensate for traditional CBF and CF to yield good recommendation results.

[1] An active user is the user for whom the recommendation prediction is made.

The rest of the paper is organized as follows. In Section 2, the relevant literature is reviewed. Section 3 presents our proposed approach. Experiments and corresponding results are shown in Section 4 to justify our proposed approach. Finally, concluding remarks, research limitations and future work are addressed in Section 5.

2 RELATED WORKS

2.1 Filtering Approach

In the mid-1990s, recommender systems have appeared to be an important research area to help users overcome information overload problem and to provide personalized recommendations (Adomavicius & Tuzhilin 2005). The purpose of recommender systems is to facilitate our information filtering process by automatically recommending desired information through the analysis of our past preferences or the preferences of other individuals who share similar interests.

Ever since the emergence of recommender systems, a variety of filtering methods have been developed. Balabanović and Shoham (1997) classified them into three categories: content-based recommendations, collaborative recommendations and hybrid approaches.

The content-based filtering (CBF) approach roots itself from information retrieval and information filtering research areas. CBF focuses on items of textual format, such as documents, news, movie reviews and Web sites, and makes recommendation by analyzing the item content to look for the commonalities among the items that the user preferred in the past. This kind of analysis will construct a user profile which includes the users' tastes, preferences and needs. Only the items that have high degree of similarities with the user profile will be retained and recommended to the user.

Unlike CBF, the alternative approach, collaborative filtering (CF), relies only on the user-item rating matrix to predict the utility of items that a particular user might like. Traditional collaborative filtering algorithms employ the entire collected user-item ratings to make the predictions based on the similarity among users (user-based CF) or items (item-based CF). User-based CF analyzes a users' group which share similar interests or common experience with the user and recommends items that this group generally prefer, while item-based CF recommends items which have high similarity with the list of items an active user had rated in the past.

The hybrid filtering approach aims at how to appropriately combine the CBF and CF to complement each other and improve the recommendation performance. Depending on how both approaches are combined, the hybrid ones can be classified into the following (Adomavicius & Tuzhilin 2005): (1) implementing CBF and CF separately, and aggregating the results using certain schemes; (2) exploring content-based features into CF similarity computation; (3) integrating collaborative user-item ratings into CBF item-feature to generate user profiles for analysis; and (4) constructing a unifying model that exploits both ratings and content information by some sort of inductive learning approaches.

2.2 Latent Topic Analysis Model

Early research works on representing a document include the vector space model in information retrieval where a document is expressed by a vector of keyword weights. However, this kind of representation scheme provides limited reduction in description length and reflects little about the intra- or inter-document structure. Further dimension reduction techniques are therefore developed to tackle these problems including probabilistic latent semantic analysis (pLSA), and latent Dirichlet allocation (LDA). Both of them aim at resolving the curse of dimensionality problem by capturing hidden semantic structure in document modeling.

Hofmann (1999) proposed the probabilistic latent semantic analysis (pLSA). pLSA assumes some underlying latent semantic structure hidden between words and documents, and defines a statistical model referred as the aspect model where each observation, the co-occurrence of a word (w) in a particular document (d), is associated with an unobserved class aspect or topic (z). As a result, each word is generated from a single topic and different words in a document may be generated from different topics. Each document is represented by the mixture topics and reduced to a probability distribution on those topics. The resulting distribution is the “reduced description” associated with the document.

Although pLSA makes a great improvement in probabilistic modeling of text documents, it is incomplete because it does not provide probabilistic model at the level of documents. This may result in such problem as linear growth of the parameters estimated in the model that tends to overfitting easily, and no natural way to assign probability to a previously unseen document.

Latent Dirichlet Allocation (LDA) was proposed by Blei et al. (2003) to compensate for pLSA. LDA is an unsupervised probability generative model that can randomly generate the documents that are observed. Particularly, it can identify “hot topics” by uncovering the temporal dynamics of latent topics in documents. It overcomes both of the problems with pLSA by treating the topic mixture weights as a K-parameter hidden random variable rather than a large set of parameters that are directly linked to the training set. The authors have shown that LDA is capable of capturing the latent semantic information from a collection of documents, and demonstrates its superiority compared to several other models which includes the multinomial mixture (MM) (Nigam et al. 2000) model and pLSA.

LDA has been applied in several fields such as text segmentation (Misra et al. 2011), tag recommendation (Krestel et al. 2009), automated essay grading (Kakkonen et al. 2005), topic identification (Griffiths & Steyvers 2004), fraud detection in telecommunications (Xing & Girolami 2007), and Web spam classification (Bíró et al. 2009). Misra et al. (2011) proposed a LDA-based approach to segmenting a text into semantically coherent parts. LDA was used to detect segment boundaries because a segment change should be associated with a significant change in the topic distribution of the segment. In addition, a modified dynamic programming algorithm was incorporated to speed up the segmentation process without loss in performance.

3 PROPOSED APPROACH

As stated, the objective of this research is to propose hybrid filtering approaches to make personalized document recommendation. Latent Dirichlet allocation (LDA) is employed to uncover the semantic structure hidden in the documents that users have read, including the word distributions over the latent topics and the latent topic mixture distribution over documents. After LDA model is established, we then incorporate the result into item-based CF similarity computation to facilitate the CF prediction process, i.e., the document similarity is measured by comparing the latent topic distributions of two documents, instead of comparing by the document vectors extracted from the rating matrix in the traditional way.

The proposed approach, referred to as semantic-based collaborative filtering (SBCF), basically consists of four steps, which are discussed in the following.

Step 1: Building LDA Model

The first step is to build LDA model from the collected documents which users have seen or liked. In LDA, latent topics are assumed to be multinomially distributed over documents (denoted by θ), and words in a document are assumed to be multinomially distributed over latent topics (denoted by ϕ). With these distributions explored, we can identify the semantics of the latent topic space by relating them to words and documents. Explaining the prominent words related to each latent topic will help us understand the nature of the topic, and explaining the prominent latent topics related to each document will help us understand the nature of the document.

In literature, variational EM (Expected Maximization) (Blei et al. 2003) and Gibbs sampling (Griffiths & Steyvers 2004) are two common approaches that can be applied for the estimation of θ and ϕ from the collected unlabeled corpus. Nevertheless, most of research works focus on Gibbs sampling since its performance is comparable to variational EM but faster in convergence and better tolerant to local optima.

Accordingly, in this study, we employ the Gibbs sampling to estimate parameters of LDA which iterates multiple times over each word v to sample a new topic k for the word based on the probability $p(z_i = k \mid v, z_{-i})$ as follows:

$$p(z_i = k \mid v_i, z_{-i}) \propto (n_{d,k} + \alpha_k) \frac{n_{k,v} + \beta_v}{\sum_{v'} n_{k,v'} + \beta_{v'}} \quad (1)$$

where $n_{v,k}$ maintains a count on topic-word assignments, $n_{d,k}$ counts the document-topic assignments, z_{-i} stands for all topic-word and document-topic assignments except the current assignment z_i for word v, and α and β are the parameters for the Dirichlet priors, serving as smoothing parameters for the counts. Through the counts of posterior probabilities in eq. (1), parameters θ and ϕ are obtained as follows:

$$\theta_{d,k} = \frac{n_{d,k} + \alpha_k}{\sum_{k'} n_{d,k'} + \alpha_{k'}} \tag{2}$$

$$\phi_{k,v} = \frac{n_{k,v} + \beta_v}{\sum_{v'} n_{k,v'} + \beta_{v'}} \tag{3}$$

Step 2: Measuring Similarity between Documents

This step is to use LDA results to find out the similarity between documents in order to facilitate item-based CF prediction. The estimated θ denotes the latent topic distribution of each document, viewed as a matrix of documents by topics, and can be applied to calculate the similarity between documents.

Since each document has its own topic distribution from θ, we then apply the cosine-based similarity measure between any two documents by viewing each of them as a (topic-based) vector. This is expressed as

$$sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \times \|\vec{j}\|} \tag{4}$$

where "·" denotes the inner-product of the two vectors.

Step 3: Expanding active user’s preferences

In this step, we desire to expand active user’s preferences based on the document similarity result with the aim to predict his/her interests toward unseen documents. Assume that we are given users’ reading records of documents as a matrix *R*, which is the rating matrix employed in CF in the usual sense. In our study, the element values of *R* are either “1”, denoting the user has seen and liked this document, or “– ”, denoting the user has not seen this document yet. We then look at the set of documents an active user has seen and determine how similar they are to other documents the active user has not seen using the similarity matrix from the previous step. By doing so, the ratings of unseen documents for the active user can be obtained and they serve to indicate the preference degrees of the active user toward the unseen documents.

For item-based CF, the predicted rating $P_{a,i}$ for a novel document *i*, with respect to an active user *a*, is based upon the weighted average of ratings from all other documents that have been rated by the active user *a*. The formula is listed in the following:

$$P_{a,i} = \sum_{n \in N_a} w_{i,n} \tag{5}$$

where $w_{i,n}$ is the weight (similarity degree) between documents *i* and *n* from the similarity matrix. The summation is taken over all rated documents $n \in N_a$ by user *a*.

Step 4: Predicting Top-N Recommendation

Finally in this step, we would like to construct the top-*N* recommendation set for the active user. Results from the previous step show the predicted preferential ratings for unseen documents to the active user. Therefore, we can simply sort the ratings in the descending order and select the first *N* documents that are of top-*N* predicted ratings to generate the recommendation list.

4 EXPERIMENTS AND RESULTS

In this section, we conduct two experiments to examine the performance of our proposed hybrid filtering approach, SBCF, as described in Section 3. In addition, we compare the performance with three other approaches: content-based filtering[2] (TFIDF), user-based CF (UBCF) and item-based CF (IBCF) that serve as baselines for comparison.

4.1 Experimental Design

In our experiments, we collect the dataset from CiteULike (http://www.citeulike.org/) which is commonly applied in document recommendation and tag recommendation. CiteULike is a website that aims at assisting to store, organize, and share scholarly papers that users are reading. Users can name the scholarly papers they are interested in with tags (bookmarks). CiteULike offers data files on a daily basis that constitutes anonymous dumps of who posted what and when the posting took place. We utilized the dumps from January 6th 2006 to January 5th 2007 as the collected dataset in our experiments. For the collected data, we filter users who read less than 20 documents in their personal profile since it is unreliable to predict the cold start users. We also filter out those documents that occur only once during the time period because the cold start items do not contribute significantly in the analysis. We therefore obtain a dataset, called CUL, with 3,201 documents read by 86 users.

Our study adopts Precision, Recall, and MAP to measure the recommender performance. Precision is the fraction of recommended items that are relevant. It is defined as the number of hits (i.e. the number of documents in the test set that also appears in the top-*N* recommended documents) divided by the

[2] This CBF is based on the comparison of document features extracted using the TFIDF weights.

number of all recommended documents. On the other hand, Recall is the fraction of relevant instances that are retrieved. It is defined as the number of hits divided by the number of documents in the test set. Finally, AP is the average of the precision scores after each correctly recommended document, which is defined by

$$AP = \frac{\sum_i precision@i \times corr_i}{N_{a,t}} \tag{6}$$

where *precision@i* is the precision at ranking *i* and $corr_i$ =1 if the document at position *i* is correctly recommended, otherwise $corr_i$ = 0. $N_{a,t}$ is the number of documents that user has read in the test set. *MAP* is the mean of average precision scores over all test users.

The evaluation scheme used in our approach is the 10-fold cross-validation where the data are divided into 10 equal-sized subsets with respect to the users. Each time, 9 of the subsets are prepared for the training and the remaining one subset is for the test. However, the actual training data contain both the 9 subsets and 50% of the remaining subset, randomly selected with respect to each user. Then the rest withheld 50% of the remaining subset is the test data to evaluate the performance. For each user in the remaining subset, we generate a top-N recommended list of documents using the training data and examine whether the withheld documents also appear in the recommended list. This procedure is repeated 10 times and the final performance is averaged over the 10 folds to obtain robust results.

4.2 Experiment I

The objective of Experiment I is to set up parameters employed in SBCF in LDA model, which includes Dirichelt hyper-parameters α and β, the Dirichlet distribution parameters, and the number of topics *Z*. In literature, some guidance is provided for these two parameters that β = 0.1 and α = 50/Z (Griffiths & Steyvers 2004). We therefore adopt this setting in our experiment.

The more difficult setting is on the number of latent topics, *Z*. It usually varies in different situations such as the selected dataset and its associated size. Blei et al. (2003) proposed a perplexity measure that is commonly applied in language modeling to evaluate the predictive power of the model. The perplexity measure for a test set of *M* documents is defined as

$$Perplexity(D_{test}) = \exp\left\{ - \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{7}$$

The lower the perplexity, the better performance the trained model will be. Therefore, by varying the number of latent topics, we can observe the trend of the perplexity measure and setup the number of latent topics when the trend reaches its minimum.

In our experiment, we use Stanford Topic Modeling Toolbox which was developed by the Stanford NLP group to build the LDA model and measure the perplexity. The result is shown in Figure that illustrates the tendency of perplexity with different number of latent topics. From this result, we do observe a U-shaped curve that reaches its minimum with the number of latent topics being 80.

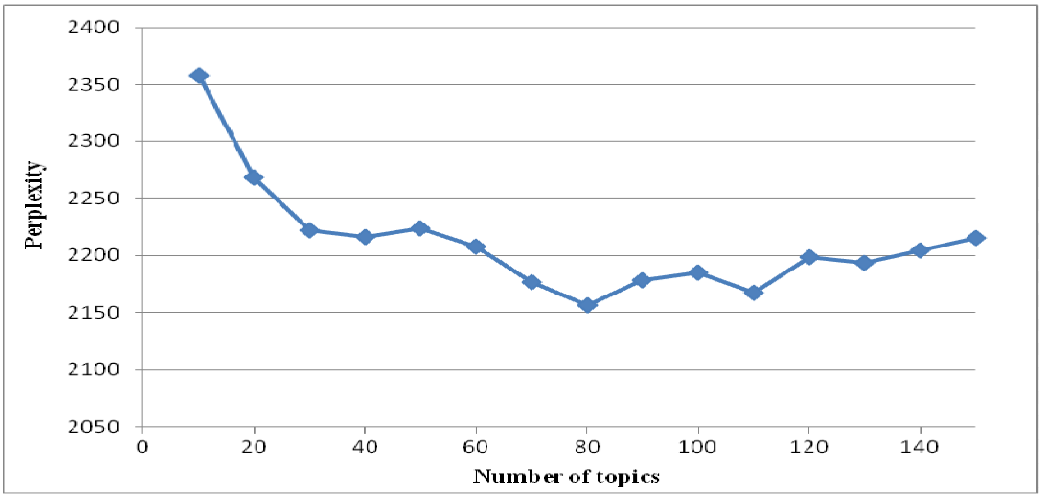

Figure 1. *Perplexity of LDA model*

However, as mentioned in literature (Asuncion et al. 2012), perplexity is not always a reliable measure to determine the number of latent topics. Therefore, in our study, we choose to determine *Z* by trial and error that varies from 10 to 300, in increment of 10 each time. The Precision performance result for SBCF is shown in Figure 2 (results for Recall is similar).

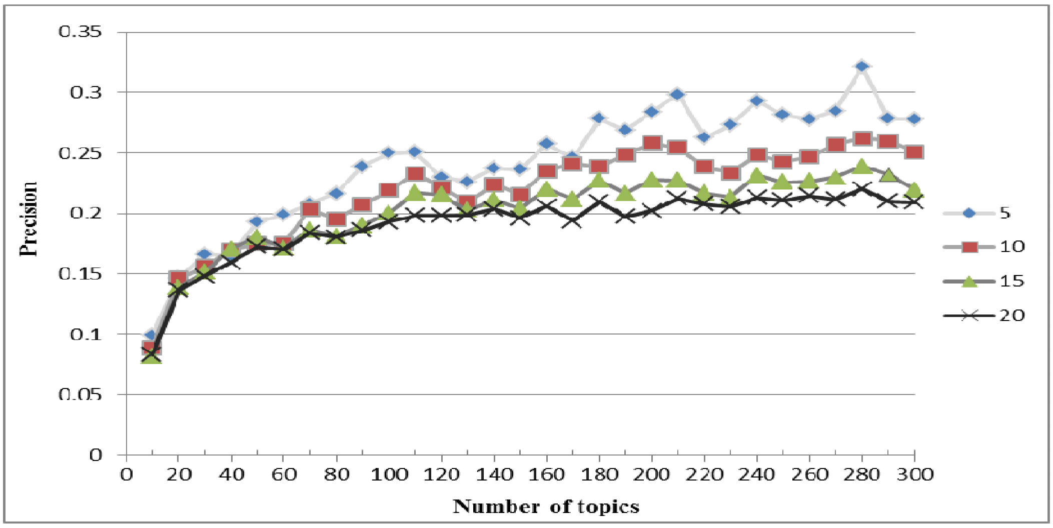

Figure 2. *Precision performance of SBCF*

From Figure 2, apparently the best performance does not occur with the number of latent topics being 80. Instead, the performance fluctuates upwards until it reaches the maximum around the number of latent topics of 280. This phenomenon is not difficult to comprehend. Latent topics serve as the similarity computation basis for document-to-document similarity in SBCF. With sufficient (not too few) and non-redundant (not too many) latent topics, both approaches will exhibit its feasibility on recommendation prediction. To summarize, we set up the parameters employed in the experiments as $\beta = 0.1$ and $\alpha = 50/Z$, and the number of latent topics, Z = 280.

4.3 Experiment II

In this experiment, we desire to compare performance of SBCF with that of TFIDF, UBCF and IBCF using CUL. Figure 3, Figure 4, and Figure 5 show the performance comparison results using *Precision*, *Recall* and *MAP*, respectively, where N denotes the top-N ratings in the recommendation list.

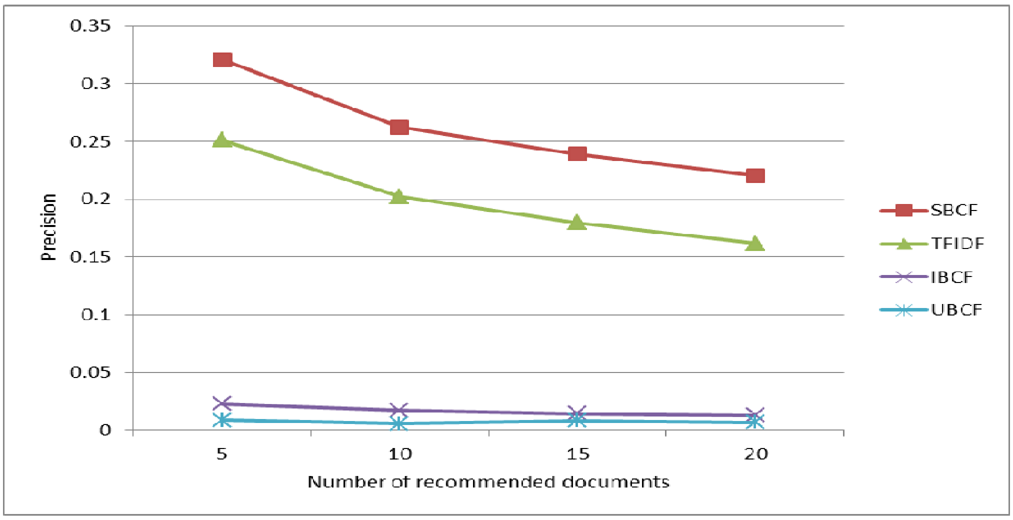

Figure 3. Comparison of Precision performance

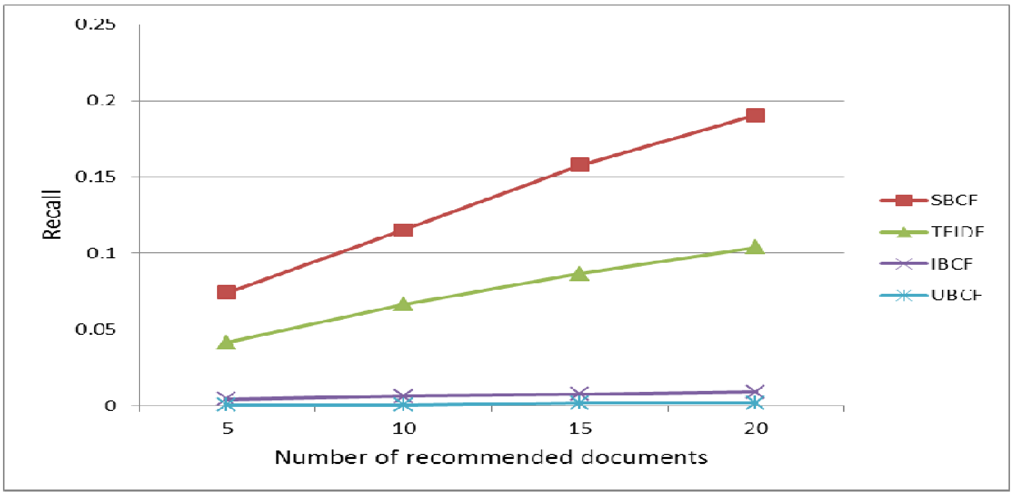

Figure 4. Comparison of Recall performance

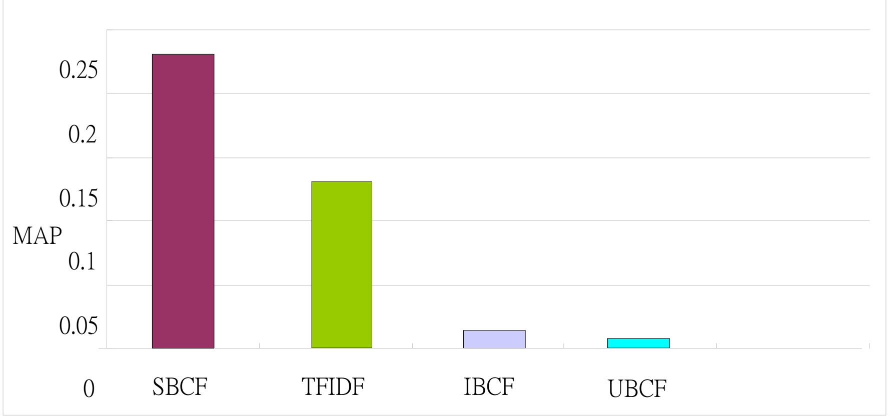

Figure 5. Comparison of MAP performance

From these figures, we first observe that both UBCF and IBCF perform worst. This is because the traditional CF approaches suffer severely from the sparsity problem in CUL[3]. This result once again reflects the unreliable predicted recommendation for pure CF approaches if the coverage of the rating matrix is highly sparse. In contrast, SBCF and TFIDF show significantly better performance compared to traditional CF approaches. TFIDF, however, is still inferior to our proposed SBCF since the traditional CBF approach may easily run into the over-specification problem that cannot expand users' preferences beyond their past profiles. This result demonstrates the necessity of employing hybrid approaches of CBF and CF for the task of personalized document recommendation, and more importantly, the LDA model incorporated into our proposed approaches exhibits its capability of performing such a task. Therefore, the feasibility of SBCF on such applications is justified.

5 CONCLUSIONS

In this research, we propose to utilize the latent Dirichlet allocation (LDA) model to analyze the latent semantic structure among collected documents before performing the filtering tasks. With LDA results, word distributions over the latent topics and the latent topic mixture distribution over documents can be uncovered. Furthermore, these results can help us to either obtain robust document similarity in CF and hence the hybrid filtering approaches, SBCF is proposed respectively in our study.

Two experiments are conducted to examine the performance of our proposed approach. The first experiment is to set up the parameters employed in SBCF such as the hyperparameters α and β in the Dirichlet distribution, and the number of latent topics Z. The second experiment is to compare SBCF with traditional content-based filtering (TFIDF), user-based CF (UBCF) and item-based CF (IBCF). The results show that SBCF and TFIDF perform much better than UBCF and IBCF under the highly

[3] The sparsity is $1 - 4870 / (86 \times 3201) = 98.23\%$.

sparse data of CUL. Furthermore, SBCF outperforms TFIDF because it does not suffer from the specification problem that limits the expansion of users' preferences beyond their past profiles. The incorporation of the LDA into the proposed hybrid approach does enhance the prediction performance significantly.

Although the results of our research seem promising, there are some issues that need to be addressed. For example, the "rating matrix" employed in our study does not conform to the usual sense in collaborative filtering because it contains no preferential ratings but only "1", indicating the document has been seen and liked by the user. To adapt our proposed approaches into more real situations, we need to collect a more appropriate dataset and examine their feasibility accordingly.

References

Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on, 17*(6), 734-749.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y.W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*.

Bíró, I., Siklósi, D., Szabó, J., and Benczúr, A.A. (2009). *Linked latent dirichlet allocation in web spam filtering.* Paper presented at the Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web.

Balabanović, M., and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM, 40*(3), 66-72.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Griffiths, T.L., and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5228-5235.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J. (2005). Comparison of dimension reduction methods for automated essay grading. *Natural Language Engineering, 1*, 1-16.

Krestel, R., Fankhauser, P., and Nejdl, W. (2009). *Latent dirichlet allocation for tag recommendation.* Paper presented at the Proceedings of the third ACM conference on Recommender systems.

Misra, H., Yvon, F., Cappé, O., and Jose, J. (2011). Text segmentation: A topic modeling perspective. *Information Processing & Management, 47*(4), 528-544.

Nigam, K., McCallum, A.K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning, 39*(2), 103-134.

Xing, D., and Girolami, M. (2007). Employing Latent Dirichlet Allocation for fraud detection in telecommunications. *Pattern Recognition Letters, 28*(13), 1727-1734.