MIS Review 22(1/2), September (2016)/March (2017), 1-20.

Hidden Topic Analysis for Personalized Document Recommendation

Te-Min Chang[1], Wen-Feng Hsiao[2], Ming-Fu Hsu[3]
[1] *Department of Information Management, National Sun Yat-sen University*
[2] *Department of Information Management, National Pingtung University*
[3] *English Program of Global Business, Chinese Culture University*

ABSTRACT: *Collaborative filtering (CF) is a common technique used by most recommender systems for reducing information overload or finding products to purchase. The limitation for applying of CF on document recommendation lies in that CF does not consider the semantics of documents, but rather collects preferential information from many users. In literature, document recommendation tasks rely more on content-based filtering (CBF) simply because document contents are handy for analysis. However, CBF methods alone have a natural limitation in the feature selection to represent the items and in serendipitous recommendations beyond items in the profile.*

The objective of this research is thus to propose hybrid filtering approaches for document recommendation system to overcome the shortcomings of each method alone. Particularly, latent Dirichlet allocation (LDA) was incorporated to uncover latent semantic structure in the collected document corpus. The hidden topic results can act as the technical bridge between CBF and CF because we can either obtain robust document similarity in CF, or to further explore user profiles in CBF. Two experiments were conducted accordingly. The results showed that our proposed approaches outperform other counterparts on the recommendation performance, which justifies the feasibility and practical applications of the proposed approaches.

KEYWORDS : *Personalized Document Recommendation, Content-Based Filtering, Collaborative Filtering, Hidden Topic Analysis, Latent Dirichlet Allocation.*

1. Introduction

With the explosive growth of information over the Internet, more and more information is disseminating and exchanging through this new channel. The large amount of information, however, results in a challenge for users to find relevant information they are interested in. This is commonly referred to as the information overload problem due to human's limited information processing ability. To alleviate this problem, many researchers have resorted to information retrieval (IR) and information filtering techniques that help practitioners develop various tools such as search engines and recommender systems to facilitate users' online information acquisition and filtration.

Among others, recommendation service has been successfully applied to support users to identify desired information by filtering undesired one. Recommender systems suggest users the relevant information through the analyses of their past preferences or the preferences of like-minded people. Accordingly, two types of filtering techniques in recommender systems have been proposed: content-based filtering (CBF) and collaborative filtering (CF). Content-based filtering techniques compare the new information with an active user's profile of past interest to predict whether he/she whom the

recommendation/prediction is for is interested in the new information, whereas collaborative filtering techniques look for collective preferences from other similar users, and recommend their common interests to the active user.

Personalized document recommendation task is essential in assisting users to locate the preferred information in terms of textual content. In literature, CBF technique is primitively employed because the document contents are handy for analysis. However, CBF easily suffers from over-specialization problem that only recommends contents highly similar to the active users' past preferences, and thus performs worse than CF. Extra information is needed to bridge the CBF and CF in personalized document recommendation tasks to enhance the resultant performance.

Recently, several hidden topic analysis approaches such as probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) have been developed. They serve as dimension reduction approaches to investigating content features by revealing hidden topics of documents from the document corpus. Such information acts perfectly as the technical bridge because we can either explore user profiles in CBF or obtain robust document similarity in CF based on the hidden topic results. Chances are that better document recommendations can be generated with the incorporation of hidden topic information.

The objective of this research is thus to propose hybrid filtering approaches for the task of personalized document recommendation. Particularly, LDA was employed to uncover the semantic structure hidden in the document corpus. The semantic structure can account for the word distributions over the latent topics and for the latent topic distributions over documents. We can utilize LDA results in such ways as exploring user profiles to further understand users' preferences, or enhancing robustness of document similarity measures. It is desired that our proposed hybrid approaches can compensate for traditional CBF or CF alone to yield better recommendation predictions.

The rest of the paper is organized as follows. In Section 2, related work is introduced, which include filtering approaches in recommender systems and hidden topic analysis models in text mining fields. Section 3 presents our proposed approaches. Experimental results and corresponding findings are shown in Section 4 to justify the proposed approaches. Finally, concluding remarks are addressed in Section 5.

2. Related work

In the mid-1990s, recommender systems that provide recommendation service 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 the information filtering process by automatically recommending desired information (e.g. books, CDs, videos, movies, and news) through the analysis of our past preferences or the preferences of other individuals who share similar interests.

When it comes to document recommendation service, the major concern is to assist users to acquire the preferred information in terms of textual content. It becomes more and more important especially under the information overload environment like the Internet where there is more than enough information disseminating and circulating around. Circumstances such as assisting researchers to decide which scholarly papers to read, supporting knowledge workers to access task-related documents to

perform tasks, or optimizing the learning environment for the learners in e-learning systems are particularly in need of the assistance of personalized document recommendations.

Due to the textual content nature, early works on document recommendations mostly rely on content-based filtering approach. New documents are recommended by matching their content with the user profile consisting of his/her preferred content features. This allows explanations of which documents to be recommended with distinct content features. For example, Mooney and Roy (2000) developed text classifiers for the task of book recommendation. However, the recommendation performance may not be satisfactory because few serendipitous recommendations beyond those documents in the profile can be generated.

In literature, collaborative filtering approach is proposed to recommend items that are beyond those in the user's profile based on other users' collective preferences. It can be further categorized into two general classes: memory-based and model-based (Adomavicius & Tuzhilin, 2005). Acknowledging the fact that documents contain latent topics, several research works attempt to address document recommendations via model-based CF approaches (Cleger-Tamayo, Fernández-Luna, & Huete, 2012; Luostarinen & Kohonen, 2013). The model-based approaches build a model (or classifier) based on the collected user-item ratings, terms, and categories. Models can be in the form of clusters, k-nearest neighbor algorithms, Bayesian networks, or probabilistic relations. Such approaches can be efficient once the model is established; however, they may not fit under the no preferential rating (read or unread) situation as they reduce to one-class classification problems, which are more difficult to confront.

However, only a few works address document recommendations via the collaborative filtering approach. The reason may lie in the rare rating information provided by the readers compared to the abundant number of words obviously embedded in documents. As a counterexample, Amazon.com employed item-based collaborative filtering to recommend books (Linden, Smith, & York, 2003) since it is relatively easy for this company to collect adequate data as the world's largest online retailer. Hess, Stein, and Schlieder (2006) proposed to integrate a document reference network and a trust network for document recommendation. The trust network is derived from trust-based collaborating filtering that utilizes extra trust information to complement original rating information.

To bridge the techniques of CBF and CF in personalized document recommendation tasks, we need extra information to proceed. Recently, hidden topic analysis such as latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) has been proposed to uncover the hidden topics of the semantic structure in the document corpus. LDA has been applied in several fields such as text segmentation (Misra et al., 2011), tag recommendation (Krestel, Fankhauser, & Nejdl, 2009), automated essay grading (Kakkonen et al., 2008), fraud detection in telecommunications (Xing & Girolami, 2007), and Web spam classification (Bíró et al., 2009). LDA can be applied in document recommendation as well because the discovered hidden topics serve perfectly as the link to incorporate CBF and CF to enhance the recommendation performance.

3. Proposed approaches

As stated, the objective of this research is to propose hybrid filtering approaches for personalized document recommendation. Latent Dirichlet allocation (LDA) is employed to uncover the semantic structure hidden in the documents. After LDA model is established, we propose two different approaches

to incorporating the LDA results into recommendation. The first one is to apply them directly into item-based CF similarity computation to facilitate the CF prediction process, i.e., the document similarity is measured by the latent topic distributions of two documents. The second approach is to explore user profiles where latent topics of each profile are revealed by applying the LDA results. The former approach is hereinafter referred to as semantic-based collaborative filtering (SBCF), and the latter approach as collaborative-based profile filtering (CBPF). Figure 1 shows the steps of both approaches, which are described in details in the following sections.

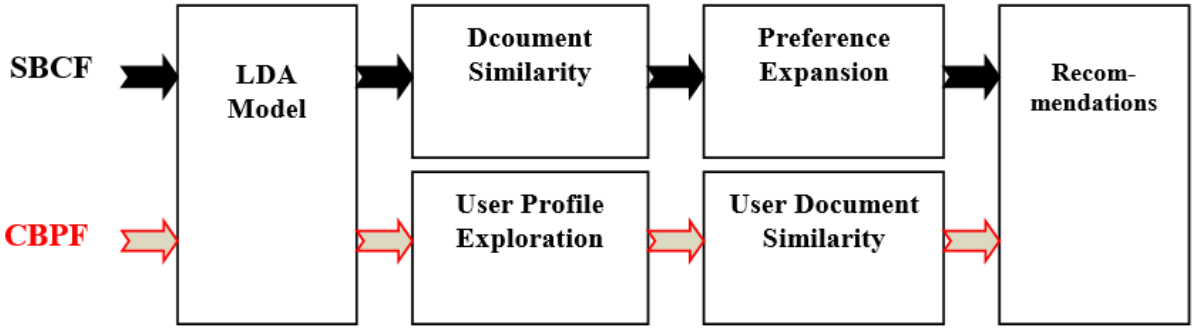

Figure 1 Steps of proposed approaches

3.1 LDA modeling

The first step of both SBCF and CBPF approaches is to build the LDA model from the collected documents which users have seen or liked. LDA is a generative probabilistic model for documents. It assumes that each document is composed of a mixture of latent topics that follow a multinominal distribution with parameters generated by Dirichlet distribution, and each latent topic is composed of a mixture of words that follow a multinominal distribution with parameters generated by Dirichlet distribution. Figure 2 shows the generative model of LDA where M denotes the number of documents (d), N denotes the number of words (w) in a document, K is the number of topics (z), α is the Dirichlet parameter specifying the document-topic distributions θ, and β is the Dirichlet parameter specifying the topic-word distribution φ.

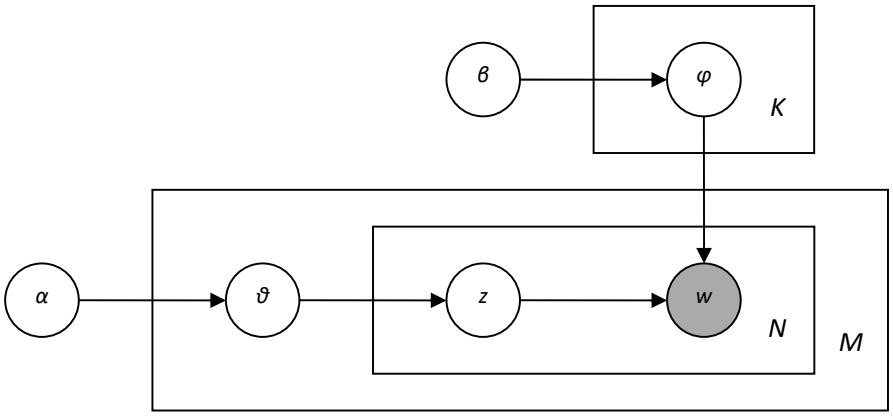

Figure 2 LDA generative model

The major task in LDA modeling is to estimate the two random variables, θ and φ, along with their associated parameters α and β, respectively, from the collected unlabeled corpus. It is shown in literature that variational EM (Expected Maximization) (Blei, Ng, & Jordan, 2003) and Gibbs sampling (Griffiths & Steyvers, 2004) are two common approaches that can be applied for the estimation. 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 (Darling, 2011).

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

$$p(z_i = k|w, z_{-i}) \propto (n_{d,k} + \alpha K) \frac{n_{k,w} + \beta w}{\sum_{w'} n_{k,w'} + \beta w} \tag{1}$$

where $n_{k,w}$ 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 w, W is the total vocabulary in the document collection, 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 can be obtained as follows:

$$\theta_{d,k} = \frac{n_{d,k} + \alpha K}{\sum_{k'} n_{k',w} + \alpha K} \tag{2}$$

$$\varphi_{k,w} = \frac{n_{k,w} + \beta W}{\sum_{w'} n_{k,w'} + \beta W} \tag{3}$$

3.2 Semantic–Based Collaborative Filtering (SBCF)

After the LDA modeling step, SBCF utilizes the hidden topic results to serve as a basis of document similarity measures. Such similarity results can be easily applied in the usual item-based CF prediction process. Consequently, SBCF goes through steps of measuring document similarity, expanding active user's preferences, and making the final recommendation predictions. These steps are further discussed in the following.

3.2.1 Measuring document similarity

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. It is viewed as a matrix of documents by topics, and can be applied to calculate the similarity between documents. Here we apply the cosine similarity as the similarity measure between any two documents and we can obtain a document similarity matrix S as a result.

For example, assume that we collect three documents and fix the number of the latent topics to be three. The LDA results for θ are further assumed as those shown in Table 1. From Table 1, the first document (D_1) is distributed over the three topics (T_1, T_2, and T_3) with probabilities of 0.2, 0.5, and 0.3, respectively. Documents D_2 and D_3 can be interpreted in a similar manner. Therefore, we can apply the similarity measure between any two documents and get a similarity matrix S. Table 2 shows the document similarity results using the data in Table 1. Note that in this simple example, D_1 is highly similar to D_2 with a similarity degree of 0.925 because both emphasize T_2 more but T_1 and T_3 less. While D_2 and D_3 do not share a lot in common with a similarity degree of 0.309 because D_3 emphasize T_1 more but T_2 less.

Table 1 Estimated θ of the LDA model

	T_1	**T_2**	**T_3**
D_1	0.2	0.5	0.3
D_2	0.1	0.75	0.15
D_3	0.7	0.1	0.2

Table 2 The document similarity matrix S

	D_1	**D_2**	**D_3**
D_1	1	0.925	0.552
D_2	0.925	1	0.309
D_3	0.552	0.309	1

3.2.2 Expanding active user's preferences

In this step, we desire to expand active user's preferences based on the document similarity result 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 typical 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 S from the previous step. In this regard, 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. In our study, however, we assume no preferential degrees in R but only "1" or "–". This is somehow limited by the data collection because in personalized document recommendation, we usually obtain what documents users have read instead of the degrees to which documents users have read and liked. This implies that traditional prediction formula for $P_{a,i}$ cannot be directly applied. We therefore modify the formula into the sum of similarity degrees between all read documents n and the unseen document i, as follows

$$P_{a,i} = \sum_{n \in N_a} w_{i,n} \quad (4)$$

where N_a is the document set that user a has read and $w_{i,n}$ is the similarity degree between documents i and n from the similarity matrix S.

Again, we take a simple example to illustrate the above processes. Assume that we obtain the users' reading records R as in

Table 3. The active user 2 has seen documents D_1 and D_2 but not D_3. To infer how user may like D_3, we simply sum up the similarity degrees of D_1 to D_3, and D_2 to D_3, which are 0.552 and 0.309, respectively from the similarity matrix S in Table 2. Then

$$P_{2,3} = 0.552 + 0.309 = 0.861 \quad (5)$$

is the expanded preference for user 2 toward unseen document D_3. The final results of the expanded preferences in this example are listed in Table 4.

Table 3 Users' reading records of matrix R

	D_1	**D_2**	**D_3**
U_1	—	—	1
U_2	1	1	—
U_3	—	1	—

Table 4 Expanded users' preferences of documents

	D_1	D_2	D_3
U_1	0.552	0.309	1
U_2	1	1	0.861
U_3	0.925	1	0.309

3.2.3 Making top-N recommendations

Finally, SBCF performs the final step of recommendation predictions based on the predicted ratings for unseen documents based on equation (4). We simply rank the ratings in a descending order and select the first N documents to generate the top-N recommendation list for each active user.

3.3 Collaborative-Based Profile Filtering (CBPF)

After the LDA modeling step, CBPF utilizes the hidden topic results to explore user profiles. The similarity between an unseen document and the user profile can then be easily measured over the latent topic distributions. Consequently, CBPF goes through steps of exploring user profiles, measuring user-document similarity, and making the final recommendation predictions. These steps are further discussed in the following.

3.3.1 Exploring user profiles

This step is to apply the LDA results to explore each user's preference over the latent topics in his/her profile. By transiting the user-document relationship (from the rating matrix R) and the document-topic relationship (the estimated θ random variables in LDA model) with average operation into the user-topic relationship, the topic distribution in a user profile can be easily inferred as follows:

$$T = a \times R \times \theta \qquad (6)$$

where a is the normalization factor for each row in T and "–" in R can be deemed as 0 under the matrix multiplication operation.

Assume again the LDA results for θ are shown in Table 1, and the users' reading records of documents of matrix R are shown in

Table 3. Note that each row of R serves as each user's reading profile. We average the topic distributions of documents in the user profile to obtain the inferred latent topic distribution for this profile. The results are shown as in Table 5. The inferred topic distribution indicates how preferable the topics interest the user.

Table 5 Topic preferences of users

	T_1	T_2	T_3
U_1	0.7	0.1	0.2
U_2	0.15	0.625	0.225
U_3	0.1	0.75	0.15

3.3.2 Measuring user-document similarity

In this step, we desire to calculate the similarity between each unseen document and the user profile in terms of latent topic distributions based on the cosine similarity measure. After the similarity computation is performed over all documents and all users, we can get a similarity matrix I as

$$I = T \times \theta^T \quad (7)$$

where θ^T is the transport matrix of θ. In fact, the similarity results for unseen documents are simply their predicted ratings.

Again, take the previous simple example. Table 1 shows how the three documents are distributed over the latent topics and Table 5 shows how the three user profiles are distributed over the latent topics. We then apply the cosine similarity measure between users and documents. Table 6 shows the similarity results for the similarity matrix I. Note that in this simple example, U_1 would fairly prefer unseen document D_1 (with 0.5519 similarity degree) and U_3 would highly prefer unseen document D_1 (with 0.9253 similarity degree).

Table 6 Similarity between documents and user profiles

	D_1	D_2	D_3
U_1	0.5519	0.3087	1.0
U_2	1.0	1.0	0.4246
U_3	0.9253	1.0	0.3087

3.3.3 Making Top-N recommendations

Finally, with the results of matrix I, CBPF can proceed in the final step to sort the ratings for the unseen documents in the descending order and select the first N documents that are of top-N predicted ratings to generate the recommendation list for each active user.

4. Experiments and results

In this section, we conduct two experiments to examine the performance of our proposed hybrid filtering approaches. 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 assists users to store, organize, and share scholarly papers. Users can annotate the scholarly papers they are interested in with tags (bookmarks). Therefore, the information provided in CiteULike fits appropriately the domain of document recommendations.

We utilize the bookmark data collected from January 1st 2012 to April 30th 2013 as our experimental data. Essentially, we only extract the document abstracts because they ease the computational efforts of LDA analysis, which may be considered as a restriction of LDA applications. Furthermore, we filter out users who read less than 20 documents in their personal profile since it is more difficult and unreliable to predict these 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 either. We therefore obtain a dataset, called CUL, consisting of 495 users, 13,029 documents, and 36,466 bookmarks.

Our study adopts *Precision*, *Recall,* and *MAP* that are commonly applied in information retrieval fields 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 number of all recommended documents, as defined by:

$$Precision = Number\ of\ hits\ /\ N \tag{8}$$

where N is the number of recommended documents. Higher Precision values indicate more preferable documents are retrieved for the users.

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, as defined by:

$$Recall = Number\ of\ hits\ /\ N_{a,t} \tag{9}$$

where $N_{a,t}$ is the number of relevant documents (in our study, it's the number of documents that user has read in the test set).

Finally, average precision (*AP*) is used for evaluation systems that return a ranked list of documents. It considers the precision scores at each ranked position of the returned documents in the list. It is defined as

$$AP = \frac{\sum_i Precision@i \times corr_i}{N_{a,t}} \tag{10}$$

where $Precision@i$ is the precision at rank i and $corr_i$ =1 if the document at position i is relevant, 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 randomly divided into 10 equal-sized subsets with respect to the users. Each time, nine of the subsets are prepared for the training and the remaining one subset is prepared 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 (as shown in the shaded area of Figure 3). Then the rest withheld 50% of the remaining subset is the test data (as shown in the blank area of Figure 3) 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 measure the performance for the test data. This procedure is repeated 9 times and the final performance is averaged over the 10 folds to obtain robust results.

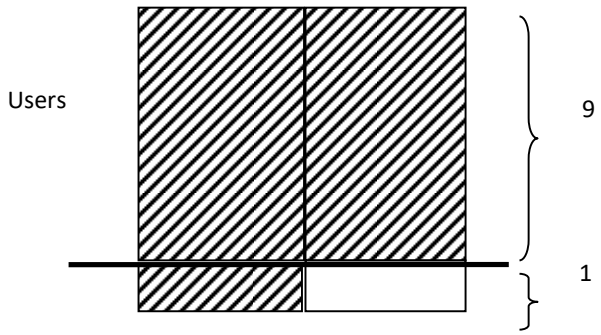

Figure 3 The schematic diagram of evaluation scheme

4.1 Experiment I

The objective of Experiment I is to set up parameters employed in SBCF and CBPF, or more precisely, in LDA model, which includes Dirichelt hyper-parameters α and β, and the number of topics K. Parameter α denotes the Dirichlet distribution parameter over the multinomial distribution of latent topics for each document, and parameter β denotes the Dirichlet distribution parameter over the multinomial distribution of words for each topic. They also serve as smoothing parameters for the counts in Gibbs sampling. In literature, some guidance is provided for these two parameters as which suggested that β = 0.1 and α = 50/K (Griffiths & Steyvers, 2004). We therefore adopt this setting in our experiment.

The more difficult setting is the number of latent topics, K. It usually varies in different situations such as the selected dataset and its associated size. We therefore select a subset out of the CUL dataset to estimate this parameter. This subset consists of 195 users, 9,616 documents, and 15,260 bookmarks. Once parameter K is estimated, it will be applied to the whole CUL dataset, as shown in the next experiment.

Blei, Ng, and Jordan (2003) proposed a perplexity measure that is commonly applied in language modeling to evaluate the predictive power of the model. 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 4 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 around 80 latent topics.

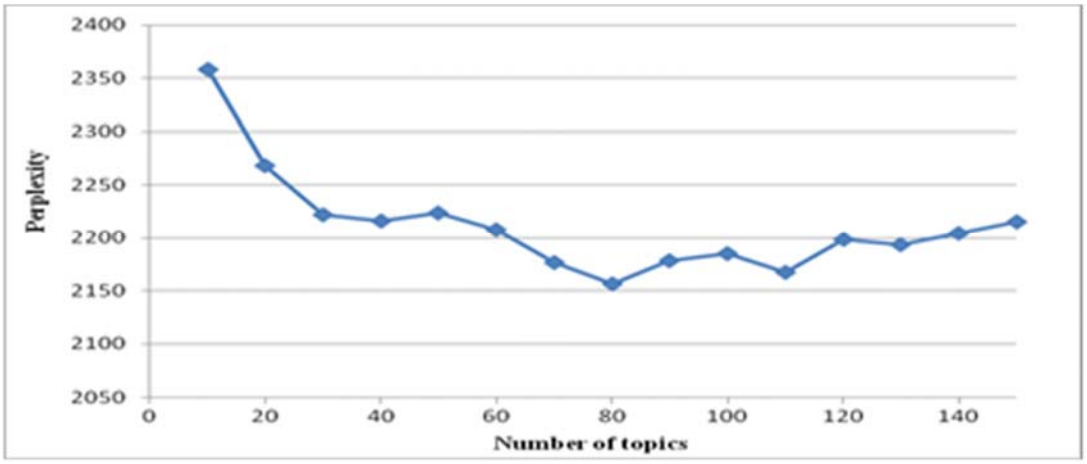

Figure 4 Perplexity of LDA model

However, as mentioned in literature (e.g. Asuncion et al, 2009), perplexity is not always a reliable measure to determine the number of latent topics. Therefore, in our study, we alternatively choose to determine *K* by trial and error that varies from 10 to 300, in increment of 10 each time. The performance results are shown in Figure 5 and Figure 6 for SBCF, and Figure 7 and Figure 8 for CBPF, respectively. They revealed the recommendation performance with different number of recommend documents N and different number of topics.

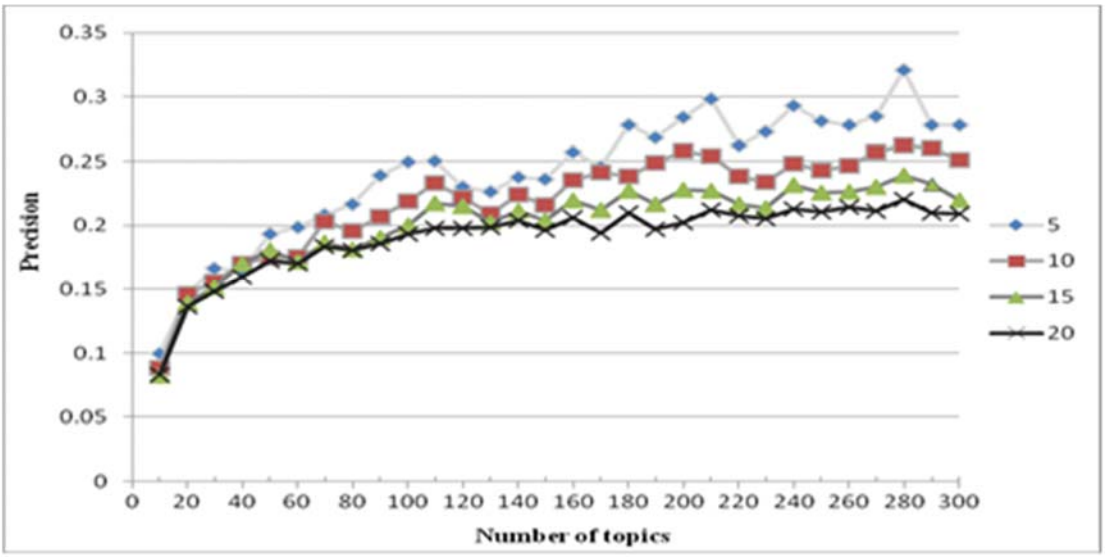

Figure 5 Precision performance of SBCF

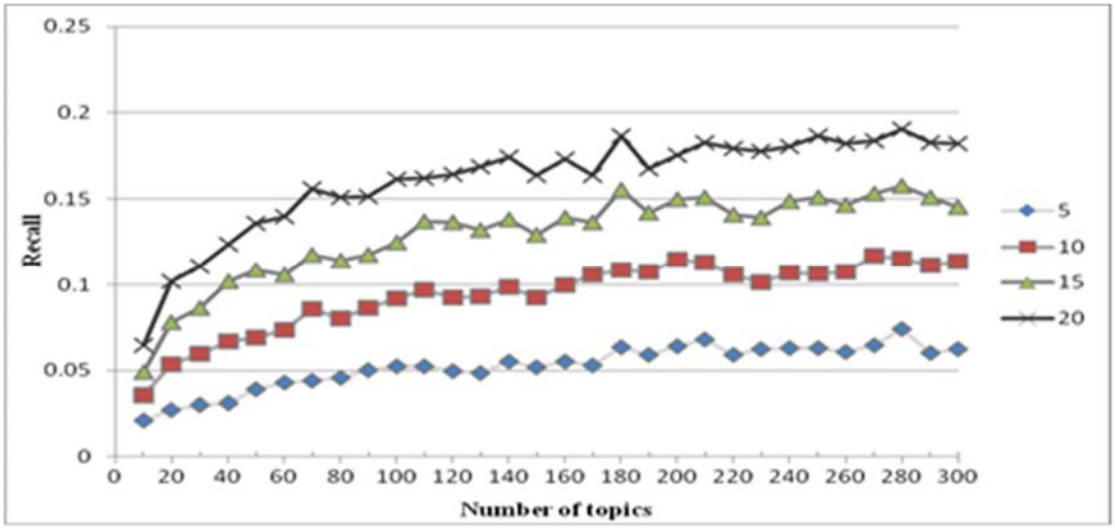

Figure 6 Recall performance of SBCF

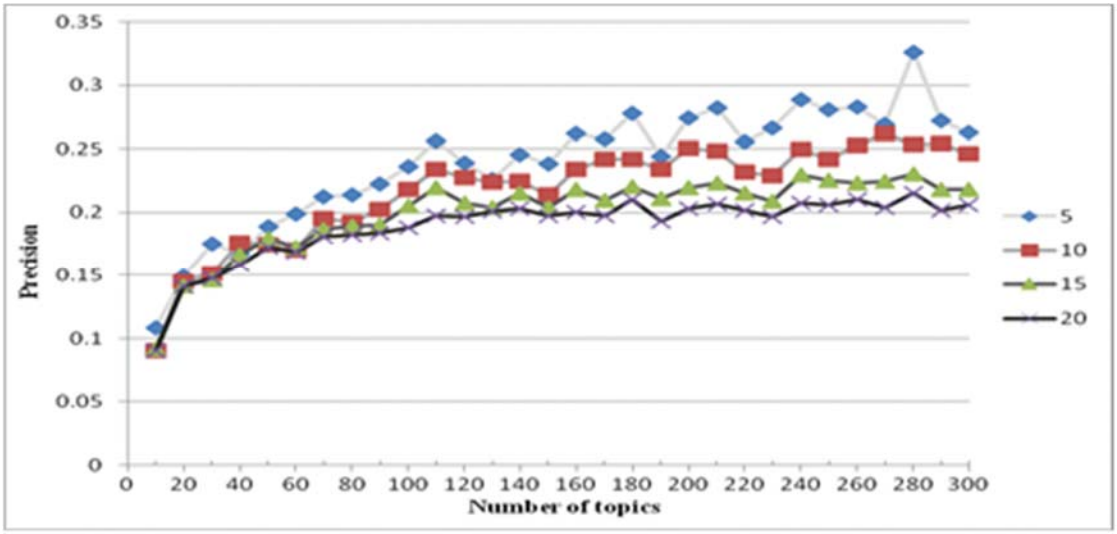

Figure 7 Precision performance of CBPF

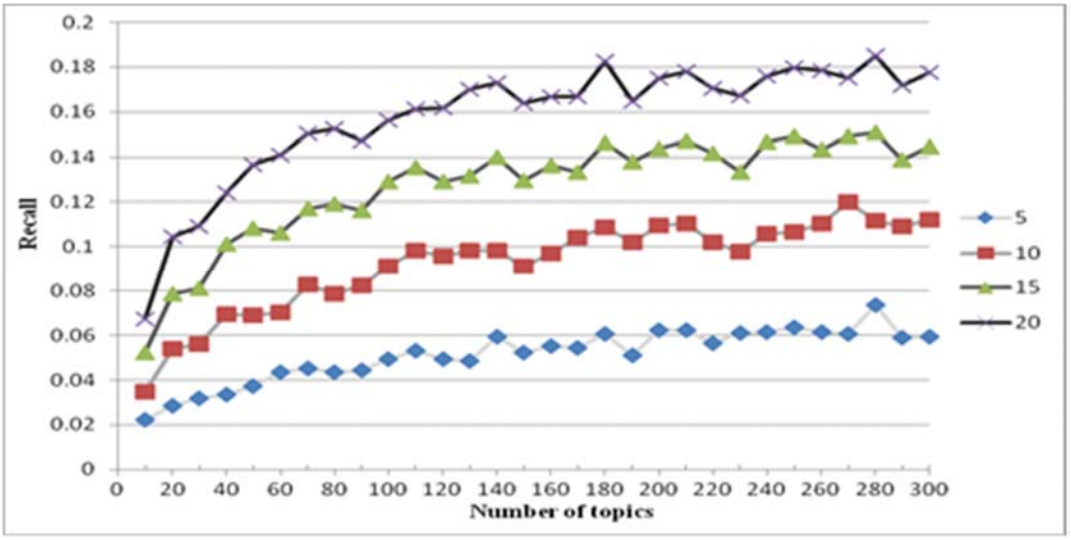

Figure 8 Recall performance of CBPF

From the above figures, apparently the best performance did not occur with the number of latent topics being 80. Instead, the performance fluctuated upwards until it reached the maximum around the number of latent topics of 280. Since latent topics serve as the similarity computation basis for document-to-document similarity in SBCF and document-to-profile similarity in CBPF, both approaches will exhibit its feasibility on recommendation prediction with sufficient (not too few) and non-redundant (not too many) latent topics. To summarize, we set up the parameters employed in the experiments as $\beta = 0.1$, $\alpha = 50/Z$, and the number of latent topics = 280 for both SBCF and CBPF.

4.2 Experiment II

In this experiment, we desire to examine the performance of SBCF and CBPF using the CUL dataset. In addition, we compare their performance with three other approaches: content-based filtering (TFIDF), user-based CF (UBCF) and item-based CF (IBCF) as baselines. For TFIDF, we extract 1000 features with the most TFIDF weights to form the document vector bases. Documents in a user profile will then be aggregated into a profiling document vector for that specific user. We compare a novel document vector with the profiling vector to determine their similarity based on which novel documents are ranked. For UBCF, we do not restrict the neighbor size. The user similarity is determined as the intersection counts of "1"s between two user ratings divided by the union counts of "1"s between them. Likewise, the item similarity in IBCF is determined as the intersection counts of "1"s between two item ratings divided by the union counts of "1"s between them.

Figure 9, Figure 10, and Figure 11 show the performance comparison results of *Precision*, *Recall* and *MAP*, respectively, where N denotes the top-N ratings in the recommendation list.

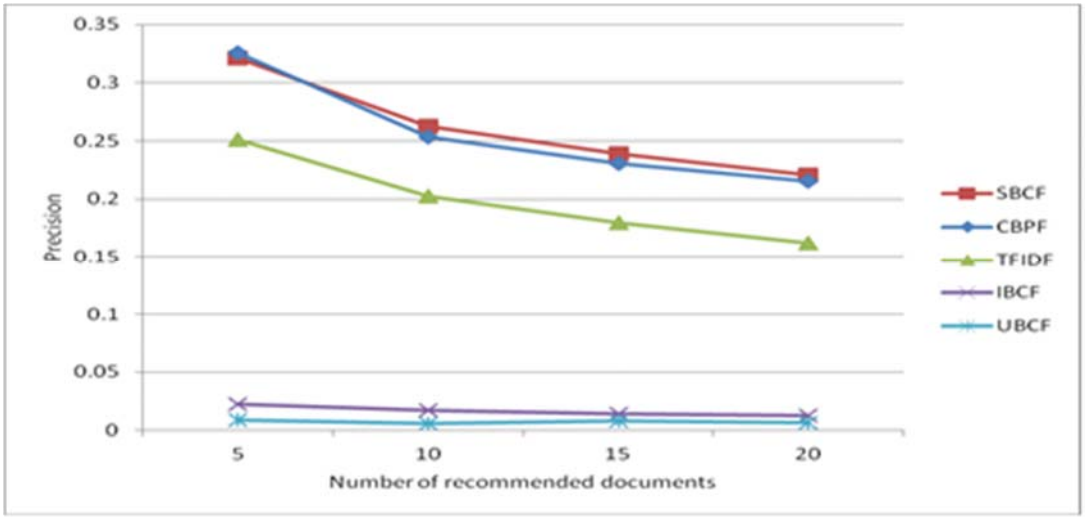

Figure 9 Comparison of Precision performance

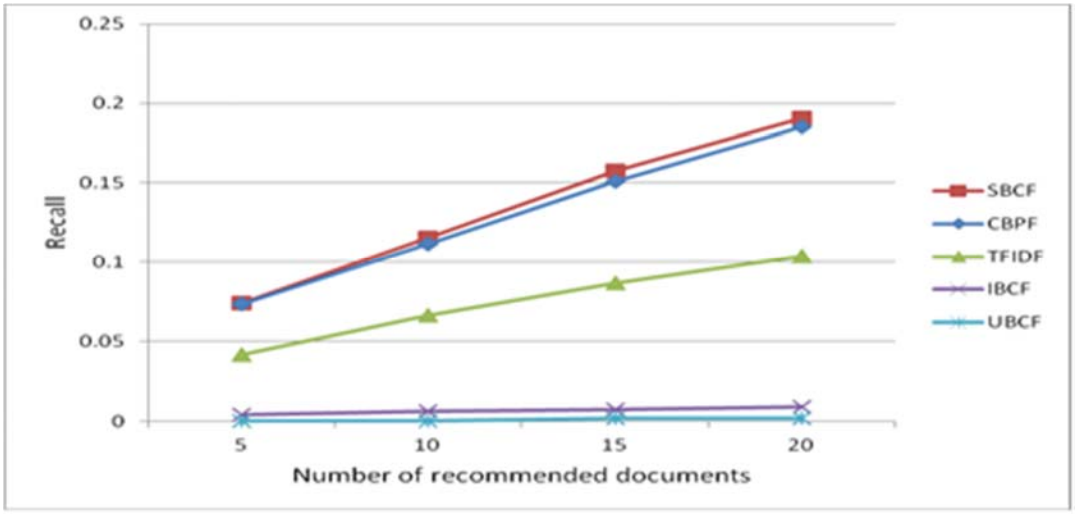

Figure 10 Comparison of Recall performance

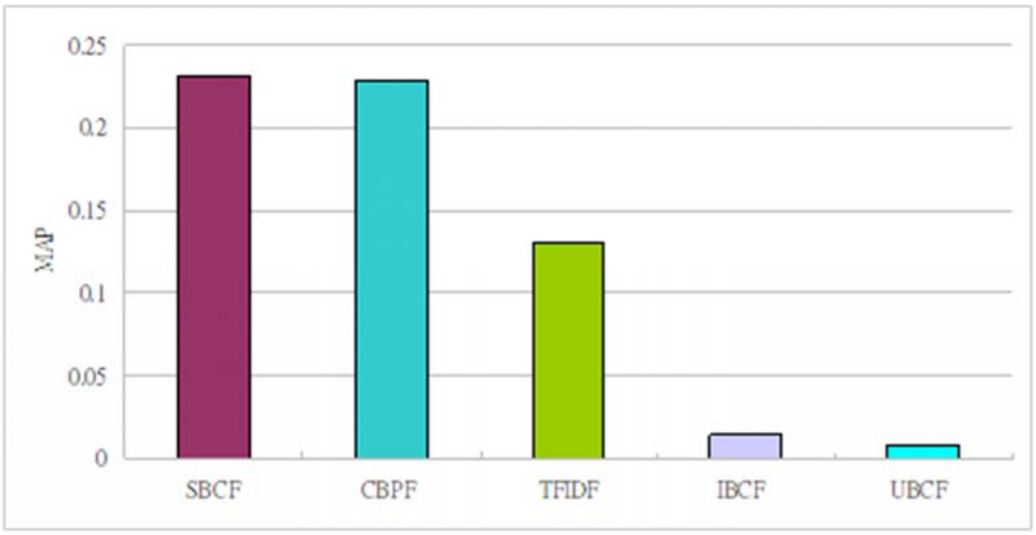

Figure 11 Comparison of MAP performance

From the above 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 (The sparsity is $1 - 4870 / (86 \times 3201) = 98.23\%$.). 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, CBPF and TFIDF show significantly better performance compared to traditional CF approaches. Personalized document recommendations may well rely on content-based filtering approaches such as TFIDF since the document content is ready for analysis. TFIDF, however, still performs below our expectation since it easily runs into the over-specification problem that cannot expand users' preferences beyond their past profiles.

On the other hand, the results indicate that SBCF does not suffer too much from sparsity because the document similarity calculation is based on the latent topic distributions over documents instead of the document ratings in the rating matrix. The results also indicate that CBPF does not suffer too much from over-specification because the user profiles have been explored based on the latent topic distributions. These outcomes demonstrate 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.

Finally, we focus on the comparison between SBCF and CBPF. Both approaches perform seemingly similarly. However, they do differ in their computational efforts. In our experiment, the average computational time is 71.39 seconds under SBCF for all test users in its Step 2 and Step 3, while it is 6.69 seconds under CBPF for all test users in its Step 2 and Step 3. The reason for this substantial difference lies in that SBCF needs to complete the computation of the entire document similarity matrix S before predicting the rating for unseen documents, and its computational time in quadratic proportional to the number of documents. As for CBPF, before predicting the rating for unseen documents, it spends time on investigating the latent topic distributions for user profiles, and its computational time is proportional to the number of documents multiplied by the number of users. In general cases as CUL, the number of

users is far less than the number of documents, and therefore, the time spends on CBPF for the predicted recommendation process will be much less than that on SBCF. To sum up, with similar prediction performance, one may still employ CBPF rather than SBCF due to the time effort for practical considerations.

5. Conclusions

In this research, we propose to utilize the latent Dirichlet allocation (LDA) model to analyze the latent semantic structure among collected documents for personalized document recommendation tasks. With LDA results, latent topic distributions over documents can be uncovered to help either obtain robust document similarity in CF, or explore user profiles in CBF. Two hybrid filtering approaches, SBCF and CBPF, are proposed accordingly 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 and CBPF such as the hyperparameters α and β in the Dirichlet distribution, and the number of latent topics. The second experiment is to compare SBCF and CBPF with traditional content-based filtering (TFIDF), user-based CF (UBCF), and item-based CF (IBCF). The results show that both SBCF and CBPF perform much better than TFIDF, UBCF, and IBCF. The incorporation of the LDA results into the proposed hybrid filtering approaches does enhance the prediction performance significantly.

Finally, the comparison between SBCF and CBPF shows insignificant different performance between them on *Precision*, *Recall* and *MAP*. However, the computational requirement for CBPF is much less than SBCF since SBCF takes time to obtain a full matrix of document similarity before recommendation prediction. Therefore, for practical consideration, one may employ CBPF rather than SBCF to perform the task of personalized document recommendation. To conclude, the experiment results do justify the feasibility of SBCF and CBPF in real applications.

Although the results of our research seem promising, there are some issues that need to be further addressed. First, the "rating matrix" employed in our study does not conform to the usual sense in collaborative filtering because it contains no preferential ratings (such as on the Likert scale) 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.

Second, although LDA exhibits its capability of enhancing the document recommendation performance, it has its own limitation of intense computation effort requirements when building the model. Therefore, in current experiments, we only utilize titles and abstracts of documents for analysis and abandon those referred documents without abstracts. However, titles and abstracts only may not be able to reflect the entire semantics in documents and cause the latent topic structure unreliable. A possible resolution to include all referred documents in LDA is to preprocess those documents to reduce their sizes while keep their original semantics. Tasks such as feature selection or text summarization may be considered to apply for the preprocessing step.

Acknowledgment

The author would like to thank Ministry of Science and Technology (MOST) of the Republic of China, Taiwan for financially supporting this work under Contracts No. 101-2410-H-110-005 and No. 105-2410-H-110-033-MY2.

References

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

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 27-34). AUAI Press.

Bíró, I., Siklósi, D., Szabó, J., & Benczúr, A. A. (2009). Linked latent dirichlet allocation in web spam filtering. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web* (pp. 37-40). ACM.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Cleger-Tamayo, S., Fernández-Luna, J. M., & Huete, J. F. (2012). Top-N news recommendations in digital newspapers. *Knowledge-Based Systems*, *27*, 180-189.

Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 642-647).

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

Hess, C., Stein, K., & Schlieder, C. (2006). Trust-enhanced visibility for personalized document recommendations. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (pp. 1865-1869). ACM.

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Journal of Educational Technology & Society*, *11*(3), 275.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems* (pp. 61-68). ACM.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76-80.

Luostarinen, T., & Kohonen, O. (2013). Using topic models in content-based news recommender systems. In *Proceedings of the 19th Nordic Conference of Computational Linguistics* (pp. 239-251).

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

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (pp. 195-204). ACM.

Xing, D., & Girolami, M. (2007). Employing latent dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, *28*(13), 1727-1734.

About the authors

Te-Min Chang*
Associate Professor
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan
Email: temin@mail.nsysu.edu.tw

Wen-Feng Hsiao
Associate Professor
Department of Information Management
National Pingtung University
Pingtung, Taiwan
Email: wfhsiao@npic.edu.tw

Ming-Fu Hsu
Associate Professor
English Program of Global Business
Chinese Culture University
Taipei, Taiwan
Email: hsumf0222@gmail.com

*Corresponding author

Te-Min Chang obtained a PhD at Department of Industrial Engineering from Purdue University, USA, in 1996. He is now an Associate Professor at the Department of Information Management, National Sun Yat-sen University, Taiwan. He has published over 20 referred journal papers including Information Sciences, Computers in Human Behaviors, International Journal of Production Research,

Annals of Operations Research, and so forth. His current research focuses on text mining, social network analysis, and decision making.

Wen-Feng Hsiao obtained a PhD at Department of Information Management from National Sun Yat-sen University, Taiwan, in 2001. He is now an Associate Professor at the Department of Information Management, National Pingtung University, Pingtung, Taiwan. He has published over 20 referred journal papers including Program: Electronic library and information systems, Journal of Information Science and Engineering, International Journal of Computer Systems Science and Engineering, International Journal of Intelligent Systems Technologies and Applications, and so forth. His current research focuses on text mining, recommendation systems, and decision making.

Ming-Fu Hsu obtained a PhD at Department of International Business Studies from National Chi Nan University, Taiwan. He is now an Assistant Professor at the English Program of Global Business, Chinese Culture University, Taiwan. He has published more than 20 journal papers including Knowledge-Based Systems, Information Sciences, Neural Computing and Applications, Economic Modelling, Connection Science, and so on. His current research focuses on financial data analysis, text mining, social network, and performance measure.

♦