LDA-BASED GROUP DOCUMENT RECOMMENDATION

Te-Min Chang
National Sun Yat-sen University, temin@mis.nsysu.edu.tw

Wen-Feng Hsiao
National Pingtung Institute of Commerce, wfhsiao@mail.npic.edu.tw

Yi-Ling Lin
National Sun Yat-sen University, yllin@mis.nsysu.edu.tw

Follow this and additional works at: http://aisel.aisnet.org/pacis2014

Recommended Citation
http://aisel.aisnet.org/pacis2014/299

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
LDA-BASED GROUP DOCUMENT RECOMMENDATION

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

Abstract

With the emergence of Internet, there is more and more information disseminating all over this channel. The abundant amount of information, however, causes difficulty for users to locate desired information, which is referred to as the information overload problem due to our limited processing ability. Therefore, recommender systems arise to assist users to acquire useful information based on their past preferences or collaborative preferences from other sources.

Group recommendation is the task of recommending items for a group of users who participate in a social activity intentionally or randomly. This kind of research is essential in the circumstances when a group of users participate in such activities as watching movies or TV shows with friends, finding restaurants for dinner with colleagues, and picking books to study in a book club. It notably distinguishes itself from personalized recommendation in that collective group behavior needs to be addressed by taking individuals’ behaviors into account.

The objective of this research is thus to propose hybrid filtering approaches for group recommendation on documents. Particularly, latent Dirichlet allocation to uncover latent semantic structure in documents is incorporated to serve as a bridge to connect content-based filtering and collaborative filtering as a whole, and generate complementary and additive effects for better performance. Two experiments are conducted accordingly. The results show that our proposed approaches (GCBPF and GSBICF) outperform other traditional group filtering approaches on the recommendation performance, which justifies the feasibility of our proposed approaches in applications.

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

With the explosive growing of the Internet, more and more information is distributed and disseminating over this channel. Nowadays, users are easily to obtain information from the Internet. It is, however, not an easy task for them to browse all the searched outcomes to acquire the desired information, which is referred to as the information overload problem due to human’s limited processing ability. To alleviate this problem, many researchers have resolved to information retrieval (IR) and information filtering (IF) techniques that help practitioners develop various tools such as search engines and recommender systems to facilitate users’ online information acquisition and filtration.

Among others, recommender systems have been successfully applied to support users to identify desired information by filtering undesired one. A typical example of the recommender system is the recommendation function provided by the YouTube video website, which could recommend videos to users by the same categories, or by analysis of users’ previous seen videos. Similarly, Amazon provides book recommendation to users through the analysis of what other users may have purchased together with those you have bought.

There are mainly three types of recommendation methods including content-based filtering (CBF), collaborative filtering (CF), and hybrid ones. Content-based filtering methods compare the novel information with an active user’s profile that records his/her past preferences to predict whether the user likes the novel information or not. In contrast, collaborative filtering methods look for similar users to the active user to identify what items are commonly interested among those users, or similar items to the novel one from the item pool to recommend. The hybrid methods simply attempt to combine both CBF and CF in different ways to complement each other and gain additive advantages.

This study focuses on the topic of group recommendations on documents. Such a research work is essential to assist a group of users to identify their desired information. It is more appropriate when a group of users participate in a social activity such as watching movies or TV shows with friends, finding restaurants for dinner with colleagues, and picking books to study in a book club. However, when it comes to group recommendations on documents, two main issues need to be addressed. The first one is about the filtering technique applied. With the nature of text content, CBF is primarily considered because we have document content ready for analysis. Unfortunately, CBF easily suffers from the problem that only recommends items highly similar to our past preferences and therefore may not perform well compared with CF. Extra information is needed to bridge the CBF and CF in document recommendation tasks to enhance the resultant performance. In this regard, we notice that hidden topic discovery techniques (e.g. Latent Dirichlet Allocation (LDA)) in the text mining field to reveal latent topics of each document can provide the required extra information to develop a better hybrid method for the document recommendation task. In other words, hidden topic information provides chances for
hybrid approach development with the aim to generate better recommendation performance.

The second issue is about the group aggregation scheme. Group document recommendation distinguishes itself from personalized document recommendation in that we need to take all individual users’ opinions into account in a form of an aggregated one to represent the group. There are several stages in the recommendation process that can perform user data aggregation into a collective data of the group and there are far more aggregation strategies that can be applied to combine individual preferences. Consequently, tasks of what stage to perform data aggregation and how to aggregate the data are involved in group recommendations.

The objective of this research is thus to propose hybrid approaches for the task of group document recommendations. Particularly, the hidden topic modeling method, LDA, is considered in our study. The hidden topic results provide good chances to explore user profile or enhance the robustness of document similarity, both of which support the development of hybrid filtering techniques. In addition, social choice theory is applied as an analytic framework of combining individual preferences, interests, or welfares to reach a collective decision.

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

2 RELATED WORK

Group recommendation is the task of recommending items for a group of users. This kind of research can be applied in the circumstances when a group of users participate in a social activity such as watching movies or TV shows with friends, finding restaurants for dinner with colleagues, and picking books to study in a book club.

In literature, group recommendation researches primarily focus on collaborative filtering techniques. Group recommendation analysis, however, distinguishes itself from personalized recommendation analysis with the need to aggregate the users’ data in a group into the collective data of the group. Therefore, the issues of what stage to perform data aggregation and how to aggregate the data need to be considered in group recommendation.

Ortega et al. (2013) pointed out that there were four stages in the CF process where the users’ data could be aggregated into the data of the group: similarity metric, neighborhood establishment, prediction phase, and recommendation list generation. According to their finding, the system performance would be significantly improved if the aggregation was done at an earlier stage of the process. This result explains why most relevant research works deal with the group aggregation strategies at the similarity
Following the above issue, we need to further consider how to aggregate the users’ data in a group at the similarity metric design stage. Masthoff (2004) employed strategies to combine individual user’s preferences into preferences of a group based on social choice theory. Social choice theory is an analytic framework of combining individual preferences, interests, or welfares to reach a collective decision (Sen, 1986). However, Arrow (1963) showed that it was simply impossible for the collective decision to satisfy all desired properties, which explains why there exist different aggregation strategies under different considerations.

Many aggregation strategies have been proposed accordingly. Masthoff (2004) listed the following possible ones: plurality voting, utilitarian strategy, Borda count, Copeland rule, approval voting, least misery strategy, most happiness strategy, average strategy, fairness strategy and most respected person strategy. Among them, approval voting scheme and least misery strategy seem to work well under the setting of our recommendation environment. The former means that each voter is allowed to vote for as many alternatives as they like, and the alternative with the most vote wins; and the latter means that each user expressed his/her rating preferences toward each alternative, and the aggregated rating of the alternative is the minimum of the individual ratings.

In addition to the above issues, group formation is another topic to address on group recommendation (Amer-Yahia et al., 2009). There are two types of group formation: intentional group and random group. For intentional group, users join the group because they share common interests on a frequent basis. For random group, users join the group without any social relations on a temporary basis such as random users joining the same cycling activity organized by a certain sports club.

3 PROPOSED APPROACHES

As stated, the objective of this research is to propose hybrid filtering approaches to making document recommendations to a group of users. With a hybrid nature, we first need to consider the group rating aggregation as in CF to deal with a group of users, instead of individuals. We then employ latent Dirichlet allocation (LDA) to further examine the document content as in CBF to uncover the latent topic mixture distribution over documents. The results can then be used to reinforce robustness of the similarity measure adopted in collaborative filtering, or advance user profile exploration in terms of latent topics.

As a result, we propose two different approaches in this study. The first approach is to apply the LDA results to further explore the user profile in terms of latent topics of documents, instead of words of documents. The similarity between a group user profile and a document can then easily computed by comparing their latent topic distributions. This approach is hereinafter referred to as group
collaborative-based profile filtering (GCBPF).

The second approach is to directly apply the LDA results into item-based CF similarity computation to enhance its robustness. That is, the item (document) similarity is measured by comparing the latent topic distributions of documents, instead of the item vectors or document vectors reading from the rating matrix. Therefore, this approach is hereinafter referred to as group semantic-based collaborative filtering (GSBCF). A conceptual framework of our proposed approaches is depicted in Figure 1. The detailed descriptions of each part are addressed in the following sections.

![Figure 1. Conceptual Framework of the proposed approaches](image)

### 3.1 Group Rating Aggregation

The first issue we consider is the aggregation scheme in group collaborative filtering techniques. As Ortega et al. (2013) pointed out that there were four stages where the individual data could be aggregated into the group data: similarity metric, neighborhood establishment, prediction phrase, and recommendation. The authors further showed that system performance would be significantly improved if the aggregation was done in an earlier stage of the process. Therefore, in this study, we adopt the same idea to obtain the aggregate data of a group from individual ones at the beginning of the CF process, i.e., the similarity metric stage.

We next consider how to aggregate the individual data into an aggregate data of a group from the given rating matrix. Several works pointed out to employ strategies based on social choice theory (e.g. Mastho, 2004; Baltrunas et al. 2010; Najjar and Wilson, 2011). Strategies based on social choice may include approval voting, utilitarian strategy, average strategy, least misery strategy, most pleasure strategy, etc.

Note that in our particular application of group document recommendation, the rating value is either “1”, indicating the user is interested in the document, or “−”, indicating the user has not read the document yet. Most of the strategies proposed under this setting cannot be applied. Therefore, we only adopt the minimal strategy where the rating of a document for the group user will be “1” if there is at least one individual user has rated “1” for the document.
3.2 LDA Model Analysis

The next issue to consider in our proposed approaches is to build LDA model from the collected documents. In LDA, each document is composed of a mixture of latent topics that follow multinomial distribution with parameters controlled by Dirichlet distribution, and each latent topic is composed of a mixture of words that follow multinomial distribution with parameters controlled by Dirichlet distribution. With the parameters fit from the given documents, we can identify the semantic structure of the latent topic space that relates documents and words.

The generative process of LDA is described as follows. First, we assign the topic-word distribution for each topic using $\beta$, and then assign the document-topic distribution for each document using $\alpha$. Finally, for each word in a document, we generate the topic using $\theta$ and choose the word using $\varphi$. Although the generative probabilistic model explains how a document generates topics and how a topic generates words, in practice we are given a collection of text documents where Dirichlet parameters $\alpha$ and $\beta$, and Dirichlet distributions $\theta$ and $\varphi$ need to be estimated. This is referred to as the inference process of LDA.

Therefore, in terms of LDA notations, we need to estimate the random variable $\theta$ and $\varphi$ that serve as the multinomial distribution parameters and are Dirichlet distributed with prior parameters $\alpha$ and $\beta$, respectively. Common practices to estimate these random variables and parameters include variational EM (Expected Maximization) (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004). In literature, however, most of research works focus on Gibbs sampling since its performance is no less than that of variational EM, but it is faster in convergence and better tolerant to local optima. Accordingly, we employ Gibbs sampling to estimate $\theta$, $\varphi$, $\alpha$ and $\beta$. We choose Stanford topic modeling toolbox to build the LDA model.

3.3 Group Collaborative-Based Profile Filtering (GCBPF)

After the issues of group rating aggregation and LDA model analysis have been addressed, we now turn to our first proposed approach, GCBPF. This approach contains user profile exploration and user document similarity as its following steps before the final recommendation predictions. These two steps are further described as follows.

User Profile Exploration

This step is to apply the LDA results to explore the user group’s preference over the latent topics in the profile. By transiting the user-document relationship (from the aggregated rating matrix) and the document-topic relationship (from the estimated $\theta$ random variables in LDA model) into the user-topic relationship, the topic distribution of the user group’s profile can be easily inferred. If the user group is interested in multiple documents, then the associated topic distributions for each specific document are averaged to obtain the ultimate relationship.
**User Document Similarity**

This step is to calculate the similarity between each unseen document and the user group’s profile in terms of latent topic distributions based on the Pearson’s correlation similarity measure. This similarity measure is exactly the predicted rating for an unseen document in GCBPF. The prediction formula is given by

$$P_{gi} = \frac{\sum_{t=1}^{K} (g_t - \bar{g})(d_{it} - \bar{d}_i)}{\sqrt{\sum_{t=1}^{K} (g_t - \bar{g})^2} \sqrt{\sum_{t=1}^{K} (d_{it} - \bar{d}_i)^2}}$$

where \( g_t \) is the group’s preference for topic \( t \), and \( d_{it} \) is \( t \)'s topic distribution in unseen document \( i \).

**3.4 Group Semantic-Based Collaborative Filtering (GSBCF)**

In this subsection, we turn to describing our second approach, GSBCF. GSBCF 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 and expanding active user’s preferences before the final recommendation predictions. They are further discussed as follows.

**Document Similarity Calculation**

This step is to use LDA results to find out the similarity between documents in order to facilitate item-based CF prediction. The estimated \( \theta \) 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. That is, we can simply take any two documents in terms of latent topics and measure their similarity by the Pearson’s correlation as

$$w_{ij} = \frac{\sum_{t=1}^{K} (d_{it} - \bar{d}_i)(d_{jt} - \bar{d}_j)}{\sqrt{\sum_{t=1}^{K} (d_{it} - \bar{d}_i)^2} \sqrt{\sum_{t=1}^{K} (d_{jt} - \bar{d}_j)^2}}$$

where \( d_{it} \) is \( t \)'s topic distribution in document \( i \).

**Group Preference Expansion**

In this step, we desire to expand the group’s preferences based on the document similarity result as its predicted ratings toward unseen documents. The similarity degree between the documents, \( w_{ij} \), obtained in equation (3-2) is a weight in the prediction formula. If an unseen document \( i \) shares sufficient similarity with those documents interested by the group, then the group ratings of those interested documents together with the weights will count. The prediction formula is given by

$$p_{gi} = \frac{\sum_{i \in L} w_{ij} \times r_{gj}}{\sum_{i \in L} w_{ij}}$$

where \( L \) is the set of interested documents for the group.
where $I_i$ denotes the neighborhood of the document $i$, $w_{ij}$ is the similarity between document $i$ and $j$, and $r_{gi}$ is group’s rating score for document $j$.

However, in our particular application, the above equation cannot be directly applied without modification since $r_{gi}$ is always equal to 1. The modified formula is simply the sum of weights, $w_{ij}$, as given by

$$p_{gi} = \sum_{j \in I_i \cap r_{gi}=1} w_{ij}$$

(4)

3.5 Recommendations

Finally, we use equation (3-1) for GCBPF or equation (3-4) for GSBCF to predict the preferential ratings of unseen documents to the user group ($P_{gi}$). We can then generate a top-N recommendation list for the group by simply sorting the ratings in a descending order and selecting the first N documents.

4 EXPERIMENTS AND RESULTS

In this section, we conduct two experiments to examine the performance of our proposed hybrid filtering approaches, GCBPF and GSBCF, as described in Section 3. In addition, we compare the performance with three other traditional approaches applying in group recommendations: group content-based filtering (GCBF), group user-based CF (GUCF), and group item-based CF (GICF) as baseline 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 provides users assistance to search, store, organize, and share scholarly papers. It offers anonymous data of who posted what and when the posting occurred. In addition, it also allows users to bookmark those papers they are interested in with tags. 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. Since there are no preferential ratings indicated, we use “1” for the bookmarked papers to represent the favorable preferences of the users; and “-” for those papers that are not present in the users’ bookmarks, which are assumed to be unseen to the users. We finally obtain a dataset, called CUL, with 495 users, 13,029 documents, and 36,466 bookmarks.

Our study adopts MAP to measure the recommender performance. $MAP$ is the mean of average precision scores (AveP) of each recommendation result, which can be shown in a finite sum as
\[
\text{AveP} = \frac{\sum_{k=1}^{N} (P(k) \times \text{rel}(k))}{\text{([relevant documents])}}
\]  

where \( \text{rel}(k) \) is an indicator function equal to 1 if the recommended document at rank \( k \) is a relevant document, 0 otherwise; and \( P(k) \) means the precision at cut-off \( k \) (i.e. only the first \( k \) documents are considered) in the top-\( N \) recommended document list. Higher MAP refers to better performance.

Next, we consider the group issues. First, as there is no indication of how users are formed in CiteULike, or put it another way, users can randomly join CiteULike, the group formation in our experiment belongs to the random group rather than intentional group. Second, in our experiments, we control the group size to be of 2, 4, or 6 users to observe the group size effect on system performance.

The evaluation scheme used in our approach is the 3-fold cross-validation where the data where the data in a group are first aggregated according to the minimal strategy described in Sec. 3.1. Then ratings of 1’s in the group data are randomly divided into 3 equal-sized subsets. Each time, two of the subsets are used for training and the remaining one is for test. We generate a top-\( N \) document recommendation list based on the training data and examine whether or not the list includes those test data as interesting (relevant) documents. This procedure is repeated 3 times for different subset setups. Finally, for each group size (2, 4, or 6), 30 groups are randomly formed. The final performance is then averaged over the (30\times3) 90 trials to obtain robust results for each group size.

### 4.2 Experiment I

The objective of Experiment I is to set up parameters that need specifying in our proposed approaches and traditional group filtering approaches under comparison. They are classified as LDA parameters, user parameter, and item parameter. To determine appropriate parameter values, we conduct several pilot studies and find out those parameters are quite consistent across different group sizes. Therefore, the following parameter determination discussions use the case of group size 2 for demonstration.

**LDA parameters**

LDA parameters include Dirichlet hyper-parameters \( \alpha \) and \( \beta \), and the number of topics \( K \). Parameter \( \alpha \) denotes the Dirichlet distribution parameter over the multinomial distribution of latent topics for each document, and parameter \( \beta \) denotes the Dirichlet distribution parameter over the multinomial distribution of words for each topic. As suggested by Griffiths and Steyvers (2004), these two parameter can be set up as \( \beta = 0.1 \) and \( \alpha = 50/K \). We therefore adopt this setting in our experiment.

The next issue is to determine the number of latent topics, \( K \). No general guidelines are provided since this number varies in different situations such as the selected dataset and its size. We therefore determine \( K \) by trial and error that varies from 50 to 500, in increment of 50 each time. The MAP performance result is shown in Figure 2 using GCBPF as an illustration.
From Figure 2, it is clear that the MAP performance fluctuates with different number of latent topics specified in LDA. The performance reaches its peak around $K = 450$. We therefore set up the number of latent topics to be 450 in all our successive experiments.

![Figure 2. Number of topics determination](image)

**User parameter**

User parameter is referred to as the number of neighbors employed in GUCF. Neighbors are contingent upon how similar\(^1\) (close) they are to the given group. We determine this parameter by trial and error that varies from 1, 2, 5, 10, 20, 50, 100, 200, and 495 (all users). The MAP performance result for GUCF is shown in Figure 3.

![Figure 3. Number of neighbors determination in GUCF](image)

From the above figure, interestingly, we notice that the best performance occurs when the number of neighbors is 1. Initially, this result is not commonly observed in user-based CF approaches. However,

---

\(^1\) Limited by the data characteristics, user similarity of GUCF is measured by (No. of interesting documents by both users / No. of interesting documents by either user).
it may be reasonable in our situation because the CUL dataset is highly sparse so that the user similarity can be unreliable. With unreliable neighbors to assist recommendations, it is better to use minimal suggestions of them. Therefore, we set up the number of neighbors to be 1 for GUCF.

**Item parameter**

Item parameter is referred to as document similarity threshold employed in GICF and GSBCF. Documents with similarity\(^2\) higher than the determined threshold are regarded as similar ones. For GICF, we vary the threshold values from 0 to 1, in increment of 0.1; and for GSBCF, we vary the threshold values from \(-1\) to 1, in increment of 0.1. The MAP performance results for GICF and GSBCF are shown in Figure 4 and Figure 5, respectively.

![Figure 4. Document similarity threshold determination in GICF](image)

![Figure 5. Document similarity threshold determination in GSBCF](image)

Figure 4 exhibits a common pattern where the MAP performance reaches its peak when the threshold is around 0.5. On the other hand, Figure 5 exhibits another interesting pattern where the MAP performance is highest and remains the same when the threshold varies from \(-1\) to \(-0.1\). This

\(^2\) Again, limited by the data characteristics, document similarity of GICF is measured by (No. of users interested in both documents / No. of users interested in either document). This similarity degree ranges from 0 to 1.
phenomenon can be explained as follows. First, negative threshold value means the identification of
dissimilar documents is useful to avoid unfavored recommendations. However, the negative document
similarity gathers around $-0.1$ to $0$ in our dataset. Therefore, if the threshold is less than $-0.1$, the
identification of dissimilar documents helps in recommendations. Once the threshold is beyond $-0.1$,
no dissimilar documents will be identified and thus deteriorate the performance. To summarize, we set
up the document similarity threshold to be $0.5$ and $-1$ (no threshold) for GICF and GSBCF,
respectively.

4.3 Experiment II

In this experiment, we desire to compare performance of GCBPF and GSBCF with that of GCBF,
GUCF and GICF using CUL. Figure 6 shows the MAP performance result under the group size of 2
(results for group size of 4 and 6 are similar).

![Performance comparison for group size of 2](image)

From Figure 6, we first observe that within the traditional group filtering approaches, GICF performs
better than GUCF, and GUCF in turn performs better than GCBF. These results confirm to literature of
filtering approaches that content-based filtering may restrict its recommendations to users’ past
preferences, and, on the other hand, collaborative filtering can enhance the performance by expanding
the users’ preferences. In addition, item-based CF is better than user-based CF because item similarity is
more reliable than user similarity. An item can be easily rated by many uses but a user may not rate many
items.

Second, it is not surprising that both our proposed approaches, GCBPF and GSBCF, perform better than
GCBF, GUCF, and GICF. With the aid of LDA analysis, we are able to explore user profile in terms of
latent topics in GCBPF, or obtain robust document similarity via the latent topic distributions in GSBCF.
This result demonstrates the necessity of employing hybrid approaches for the task of group
recommendation, and more importantly, the LDA model incorporated into our proposed approaches.
exhibits its capability of performing such a task.

Finally, we examine the group size effect on the performance of GCBPF and GSBCF in the random group case. Figure 7 shows the corresponding results. We notice that both performance increases as the group size increases. This is because when more users are involved in a group, their overall preferences becomes broader and thus ease more on the recommendation tasks. However, we also observe that with smaller group size, GCBPF performs better than GSBCF, and with larger group size, GSBCF outperforms GCBPF. The reason lies in that with more users in a group, the user-topic distributions used in GCBPF tend to average out the topic preferences, and therefore the prediction between an unseen document and the group may be inaccurate (whereas the increasing trend results from broader group preferences). On the other hand, document similarity used in GSBCF does not depend on the group size, and thus its performance can be better and better with broader group preferences.

![Figure 7](image)

Figure 7. Group size effect

5 CONCLUSIONS

In this research, we propose hybrid approaches for the task of group recommendation on documents. Particularly, the hidden topic modeling method, LDA is considered to serve as a bridge to connect CBF and CF as a whole. Consequently, three different approaches are proposed. GCBPF is to apply the LDA results to further explore the group profile in terms of latent topics of documents while GSBCF is to directly apply the LDA results into CF user similarity and item similarity, respectively, to enhance the robustness of the results.

Two experiments are conducted to examine the performance of our proposed approaches. Experiment I is to set up parameters in our proposed approaches and traditional group filtering approaches under comparison. Experiment II is to compare performance resulted from different approaches. Generally speaking, the proposed GCBPF and GSBCF approaches outperform the traditional one no matter what
group size is. This result demonstrates that with the aid of LDA analysis, we are able to explore user profile in GCBPF, or obtain robust document similarity in GSBCF.

Finally, we observe the group size effect on the performance of GCBPF and GSBCF. GCBPF performs better than GSBCF when the group size is small, but GSBCF outperforms GCBPF when the group size increases. This phenomenon is explained by the average-out impact from the latent topic distributions of multiple users in GCBPF.

Although the results of our research seem promising, there are several issues that need to be addressed. First, with the limitation of our data characteristics, the rating matrix only contains elements of “1” and “-”, instead of the ordinal preferential ratings of the Likert scale. To adapt our proposed approaches into more real situations, we need to collect other appropriate datasets with both preferential ratings and text content to examine the feasibility of our proposed approaches under such cases accordingly.

Second, with respect to the CUL dataset, the group formation is random group. However, most of group recommendation may base itself on intentional groups, particularly under the social network environments. Even though we do not expect significantly different results will be exhibited by applying our proposed approaches on intentional groups, we still need sufficient evidence to verify such a hypothesis.
References


