Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2017 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 7-19-2017

# Subject-relevant Document Recommendation: A Reference Topic-Based Approach

Yen-Hsien Lee
*National Chiayi Univ. Chiayi, Taiwan*, yhlee@mail.ncyu.edu.tw

Ya-Han Hu
*National Chung Cheng University*, yahan.hu@mis.ccu.edu.tw

Wan-Chih Hsieh
*National Chung Cheng University*, memory790225@gmail.com

Pei-Ju Lee
*National Chung Cheng University*, pjlee@mis.ccu.edu.tw

# Subject-relevant Document Recommendation: A Reference Topic-Based Approach

*Completed Research Paper*

**Yen-Hsien Lee**
Department of Management Information Systems
National Chiayi University
Chiayi, Taiwan
yhlee@mail.ncyu.edu.tw

**Ya-Han Hu**
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
yahan.hu@mis.ccu.edu.tw

**Wan-Chih Hsieh**
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
memory790225@gmail.com

**Pei-Ju Lee**
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
pjlee@mis.ccu.edu.tw

## Abstract

*Knowledge-intensive workers, such as academic researchers, medical professionals or patent engineers, have a demanding need of searching information relevant to their work. Content-based recommender system (CBRS) makes recommendation by analyzing similarity of textual contents between documents and users' preferences. Although content-based filtering has been one of the promising approaches to document recommendations, it encounters the over-specialization problem. CBRS tends to recommend documents that are similar to what have been in user's preference profile. Rationally, citations in an article represent the intellectual/affective balance of the individual interpretation in time and domain understanding. A cited article shall be associated with and may reflect the subject domain of its citing articles. Our study addresses the over-specialization problem to support the information needs of researchers. We propose a Reference Topic-based Document Recommendation (RTDR) technique, which exploits the citation information of a focal user's preferred documents and thereby recommends documents that are relevant to the subject domain of his or her preference. Our primary evaluation results suggest the outperformance of the proposed RTDR to the benchmarks.*

**Keywords:** Recommender systems, document recommendation, subject-relevant recommendation, topic-based recommendation, content-based recommendation.

# Introduction

Fast development of information technology has accelerated the growth and spread of information. Sheer volumes of information have led people to difficulties in information search and management (Lee & Lee, 2004). Recommender system (RS), a subclass of information filtering system, is a well-known technique to predict what the target users might be interested in by analyzing their profile (Zhen et al., 2010). Previous studies have shown that appropriate personalized recommendations can lower users' information overload (Liang et al. 2007; Liang et al. 2008). In real-life applications, many e-commerce websites, such as Amazon.com, eBay, CDNOW, Moviefinder, and etc., have applied RSs to analyze important product attributes and customer characteristics (Linden et al. 2003; Liu et al. 2007; Polat & Du 2008; Senecal & Nantel 2004; Wei et al. 2010; Wei et al. 2002; Xiao & Benbasat 2007; Yang et al. 2010; Chen et al. 2004). Previous studies indicated that RSs could effectively provide recommendations appealing to customers to enhance sale volumes for businesses (Chen et al., 2004; Cao & Li, 2007; Choi & Ahn, 2011; Ampazis, 2008).

Recent studies have focused on the recommendation of textual information, such as news and article recommendations (Konstan et al. 1997; Lang 1995; Phelan et al. 2009; Pazzani & Billsus 1997; Semeraro et al. 2007; Liang et al 2008; Ku et al. 2012). Content-based filtering (CBF) is a kind of recommendation approaches that makes recommendations on the basis of textual features extracted from documents (Alspector et al., 1998; Liang et al., 2007). It assumes customers are interested in items that share similar attribute values with their previously stated or observed preferences (Balabanovic & Shoham, 1997; Herlocker et al., 1999; Wei et al., 2002). Specifically, CBF makes recommendations by analyzing the similarity between the target user's profile and the item's textual content (Mooney and Roy, 2000; Balabanovic & Shoham, 1997; Herlocker et al., 1999). Usually, the target user's profile is represented as a set of terms, typically the informative words, by analyzing the items that he/she has reviewed. Similarly, the items to be evaluated are also represented using the same set of terms with different weights. CBF has been applied in various recommendation scenarios involving textual documents such as books, Web sites, and news media (Cheng & Hu, 2007).

Knowledge-intensive workers such as academic researchers, who must seek information relevant to their research topics or questions confronted in different research stages (Kuhlthau, 1993; Vakkari, 2003), have demanding needs in searching academic and publishers' libraries. However, the quantity of academic articles is increasing at an accelerating pace. A total of over 50 million academic articles have been published from 1665 to 2009 (Jinha, 2010). Previous studies indicate that the information needs of researchers would vary with the problems or difficulties they face in different stages (Kuhlthau 1993; Wang & Sogergel, 1998; Vakkari, 2003). Usually, general information related to the subjects or questions that they are exploring or understanding are required and it becomes more specific as approaching the end of their task. Wilson (1973) proposed the concept of situational relevance in information retrieval. Situationally relevant items of information are those that answer, or logically help to answer user's questions of concern. Specifically, it is assumed that all the possible answers to a question constitute the user's concern sets. An item of information is directly relevant situationally if it is a member of a concern set; or indirectly relevant situationally if it is relevant but not a member of a concern set (Wilson, 1973). Furthermore, she argued that users might prefer information, which could change personal knowledge or perception status when searching information online (Huang, 1997). As a summary, an article recommender system shall recommend documents not only similar but also pertinent to the subjects of focal users' preference to satisfy their information needs.

To address the needs for recommending documents pertinent to users' preferences or to their tasks at hand, previous studies have proposed task-focused document (or literature) recommendation technique, which analyzes focal user's task profile (e.g., usage log) for making recommendations (Mobasher et al. 2000; Srivastava et al. 2000; Hwang & Chuang 2004). Moreover, some research proposed a semantic-expansion approach, which takes keywords as document concepts and expanded them by a semantic network in order to recommend documents that are semantically relevant to a focal user's preference (Liang et al., 2008; Ku et al., 2012). Nevertheless, most of them make recommendations according to content similarities between documents and a focal user's profile. As a result, such recommendations could be over-specialized that recommended documents similar to those in the focal user's profile (Shardanand & Maes 1995; Hwang et al. 2010). Though Herlocker and konston (2001) proposed a content-independent task-focused recommendation approach, it made recommendation on the basis of item associations constructed using users' interest ratings (Herlocker & konston 2001).

The citations in the article represent the intellectual/affective balance of the individual interpretation in time and domain understanding (Cronin 1984). The presence of a citation may signify that an author has been influenced by the work of another author, but it cannot, on its own, say anything about the extent or strength of the influence (Martyn 1964). Furthermore, the sum of citations to a certain paper, author or journal from a representative sample offers an acceptable surrogate of that paper's, author's, or journal's influence on a corresponding research subject or field (Culnan 1986). In brief, a cited article shall be associated with and reflect the subject domain of the citing articles. To address the over-specialization problem of CBF approach in recommending documents, our study intends to exploit citation information pertaining to a focal user's document preferences or task profile for making personalized document recommendations. First of all, our study analyzes the contents of references in documents of interest as a whole. Generally, the citations in an article could be in a large number and might cover various subject domains relevant to the article. We therefore propose Reference Topic-based Document Recommendation (RTDR) technique to discover latent topics inherent in the references for recommending articles relevant to focal user's preference or task.

In this line, our study will address two research questions:

- Can the proposed RTDR improve performance of research article recommendations?
- Can reference information help accurately identify useful research articles for researchers?

The remainder of this paper is organized as follows. Section 2 reviews the prior research relevant to our study. Section 3 depicts the overall process of our proposed RTDR technique. Section 4 describes the experimental design and the evaluation results of the proposed RTDR technique. Finally, we conclude the study and provide some research findings in Section 5.

# Literature Review

## *Recommender Systems and Document Recommendation*

Recommender system makes recommendations by sifting through a vast collection of services or items to identify those that appear relevant or interesting to focal customers. Several recommendation approaches have been proposed and classified according to different characteristics in literature (Resnick & Varian 1997; Pazzani 1999; Schafer et al. 1999; Burke 2002; Wei et al. 2002; Beyah et al. 2003). For example, Wei et al. (2002) observed the type of data and the recommendation approaches and accordingly divided RSs into six categories, including popularity-based, content-based, collaborative-filtering, association-based, demographics-based, and reputation-based RSs.

Collaborative filtering (CF) and Content-based filtering (CBF) approaches are commonly used to develop RSs (Basilico and Hofmann 2004). The CF approach makes recommendations on the basis of the preferences of a referent user group rather than essential features or attributes of items that the focal users have favored or not. That is, CF approach associates a focal user with other users whose preferences are highly similar and utilizes the collective preferences of his or her referent user group (without referring to the contents of items) to make appropriate recommendations (Billsus and Pazzani 1998; Breese et al. 1998). Specifically, CF approach first identifies a set of "nearest neighbors" whose known preferences significantly correlate with those of the focal user using a specific similarity function. Preference to an item can then be measured for the focal user using the known preferences of these nearest neighbors (Herlocker et al. 1999). The CF approach delivers personalized recommendations, using the preferences of other users, and provides several advantages that are not offered by the CBF approach (Balabanovic & Shoham 1997; Herlocker et al. 1999). For example, by analyzing other users' preferences rather than item features, it can be appropriate for recommending items whose contents cannot be processed automatically. Furthermore, it is capable of recommending items on the basis of quality and taste. However, CF approach encounters some problems in real world situations. In practice, users may have rated only a few items, resulting in a highly sparse user-preference matrix. CF approach will be ineffective to users who do not have a sufficient number of co-rated items with other users. Besides, CF approach also suffers the cold-start problems, where items that have not been rated by a sufficient number of users cannot be effectively recommended. As a result, CF approach has an inherent tendency of recommending popular items (Mooney and Roy 2000). Finally, CF approach usually faces the scalability problem because of the pair-wise user similarity measure for identifying most similar neighbors.

CBF approach analyzes the essential features or attributes of a focal user's preference items rather than the preferences of his/her referent group. Therefore, CBF approach recommends to a focal user

the items highly relevant or similar to those he/she previously purchased or showed interests (Balabanovic & Shoham 1997; Herlocker et al. 1999). Given a set of items with known preference classes as the training examples, CBF approach can be supported by supervised classification learning algorithms to construct the personalized recommendation model. The general process of the CBF approach consists of feature extraction and selection, representation, recommendation model learning, and recommendation generation (Adomavicius & Tuzhilin 2005; Wei et al. 2002). In general, CBF assumes that there are important associations among products which can be analyzed, measured, and compared, according to their respective content attribute values (Alspector et al. 1998; Liang et al. 2007); as a result, it has been widely used in textual content recommendations, such as books, Web sites, and news media (Cheng & Hu 2007). The underlying rationale is that customers might be interested in the document if it has attribute values similar to their previously stated or observed preferences (Balabanovic & Shoham 1997; Herlocker et al. 1999; Wei et al. 2002). Examples of content-based RSs include Syskill & Webert for recommending Web pages (Pazzani & Billsus 1997), NewsWeeder for recommending news-group messages (Lang 1995), and InformationFinder for recommending textual documents (Krulwich & Burkey 1996).

However, traditional CFB approach, making recommendations by assessing the similarity of content attributes between documents, may encounter the over-specialization problem. That is, the documents recommended are restricted to those that are similar to focal user's profile (Shardanand & Maes 1995; Hwang et al. 2010). Previous research proposed a semantic-based approach to address the limitations of CBF approach (Middleton et al. 2009). Instead of keyword matching, semantic-based approach measures the similarity between documents by their semantic meanings and recommends documents that are semantically similar to the focal user's profile (Liang et al. 2008; Ku et al. 2012). Though semantic-based approach expands document concepts by its content (e.g., keywords), it does not address well the over-specialization problem. Moreover, Herlocker and konston (2001) proposed a content-independent task-focused recommendation approach, which makes recommendation on the basis of item associations constructed using existing users' interest ratings. Hwang et al. proposed a co-authorship network-based task-focused approach that measures the similarity of co-authorship between documents to identify the recommendable documents. Though the content-independent approach can alleviate the over-specialization problem, the documents they recommend may not be appropriate because the recommendation is made on the basis of other users' preferences.

## *Approaches to Topic Discovery*

In this section, we review the approaches, including the traditional document clustering and Latent Dirichlet Allocation (LDA) that are adopted for discovering topics among references in our study.

### Traditional Document Clustering

Document clustering groups similar documents into distinct clusters by analyzing document contents. A document in a resulting cluster exhibits maximal similarity to the documents in the same cluster and shares minimal similarity with those in other clusters. Most document clustering techniques emphasize document contents analysis and typically consist of three phases: feature extraction and selection, document representation, and clustering (Wei et al. 2006).

Feature extraction starts with document parsing to produce a set of features (e.g., nouns and noun phrases), excluding pre-specified non-semantic-bearing words; i.e., stopwords. Representative features are then selected from the extracted features. Feature selection is critical to clustering effectiveness and efficiency because it reduces the number of the extracted features and removes the potential biases existing in the original (untrimmed) feature set (Dumais et al. 1998; Roussinov & Chen 1999). Common feature selection metrics include term frequency (TF), term frequency and inverse document frequency (TF×IDF), and their hybrids (Boley et al., 1999; Larsen & Aone, 1999). The subsequent document representation phase chooses the $k$ features that have the highest selection scores to represent each document. As a result, each document (in the corpus) is represented by a feature vector and jointly defined by the $k$ features selected. A review of previous research suggests several salient feature representation methods, including binary (i.e., presence versus absence of a feature in a document), within-document TF, and TF×IDF (Larson & Aone, 1999; Roussinov & Chen, 1999; Wei et al., 2006). In the final clustering phase, the source documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common clustering approaches include partitioning-based (Boley et al., 1999; Cutting et al., 1992; Larson &

Aone, 1999; Spangler et al., 2003), hierarchical (Roussinov & Chen, 1999; Wei et al. 2006), and Kohonen neural networks (Guerrero et al., 2002; Roussinov & Chen, 1999).

### Latent Dirichlet Allocation

To address the problem of document retrieval through a keyword search, prior research has developed probabilistic topic modeling algorithm to discover and annotate large archives of documents with thematic information. Then, users can identify the theme they are interested in and accordingly examine and retrieve the documents of interest. The goal of topic modeling is to automatically discover the hidden structure of topics from a collection of observed documents. Topic modeling algorithms generally are statistical methods that analyze documents words to discover the themes within them, associations among those themes, and their evolutions over time (Blei, 2012).

Latent Dirichlet Allocation (LDA), the simplest algorithm of topic modeling, is proposed by Blei et al. (2003) to discover themes existing in the document collection. The basic assumption of LDA is that documents exhibit multiple topics, each of which is defined as a distribution over a fixed vocabulary. Besides, it also assumes that the topics are generated prior to the documents (or any data). LDA is said to be simple because it follows a two-stage generative process to generate the words for each document. First, it randomly chooses a distribution over topics. Secondly, for each word in a document, it randomly chooses a topic from the distribution over topics, and then randomly chooses a word from the corresponding distribution over the vocabulary. In another word, each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics. In LDA, all the documents in the collection share the same set of topics, but each document exhibits those topics in different proportion.

LDA is one kind of probabilistic modeling, which treats data as arising from a generative process that includes hidden variables. This generative process defines a joint probability distribution over both observed and hidden random variables. In LDA, the observed variables are the words of the documents; the hidden variables are the topic structure; and the task of inferring the hidden topic structure from the documents is to compute the posterior distribution of the conditional distribution of the hidden variables given to the documents, which is defined as follows:

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D} \mid W_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D})}{p(W_{1:D})}.$$

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})).$$

where $p(W_{1:D})$ is the probability of seeing the observed corpus under any topic model, $\beta_{1:K}$ are topics, $\beta_i$ is a distribution of topic $i$ over the vocabulary, $\theta_d$ is the topic proportion in document $d$, $\theta_{d,k}$ is the topic proportion of topic $k$ in document $d$, $z_{d,n}$ is the topic assignments for the word $n$ in document $d$, $w_{d,n}$ is the observed word $n$ in document $d$, an element from the fixed vocabulary.

Theoretically, the number of possible topic structures is exponentially large, and thus, the posterior is not able to compute. Modern probabilistic modeling research has developed efficient methods to approximate the posterior that can be categorized into sampling-based and variational algorithms. Sampling-based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is Gibbs sampling, where we construct a Markov chain whose limiting distribution is the posterior (Steyvers & Griffiths, 2006). In contrast, variational methods are a deterministic alternative to sampling-based algorithms. Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior (Blei et al., 2003).

## Design of Reference Topic-based Document Recommendation (RTDR) Technique

For making document recommendation, most of the content-based recommendation techniques measure the similarity between documents and user's preference, on the basis of their titles, abstracts, and/or keywords. Therefore, the documents recommended by the content-based approach are usually restricted to what the focal user has read or known; that is, the over-specialization problem. To well understand the task (e.g., research issues, treatments for a diagnosed condition of disease, and etc.)

they faced, researchers usually have to go through the literature on the domain of task they are working on. Therefore, we propose RTDR technique by exploiting the reference information to recommend subject-relevant documents. Specifically, RTDR takes the references relevant to the focal user's documents of interest as a basis and analyze the latent topics within them for assessing whether documents are suitable to be recommended. A document is worthy being recommended if it discusses the important reference topics. As shown in Figure 1, the process of RTDR technique comprises three phases, including (1) feature extraction; (2) topic discovery; and (3) recommendation. In the following, we detail the design of each phase in the proposed RTDR technique.

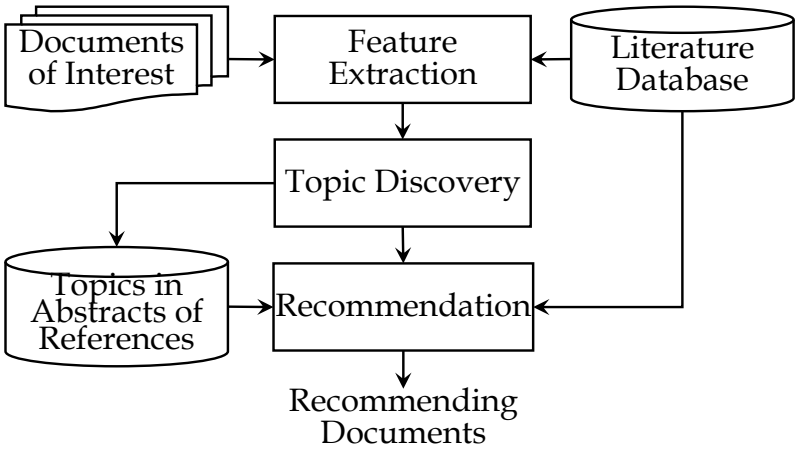


**Figure 1. Overall Process of RTDR Technique**

## *Feature Extraction*

The purpose of feature extraction is to extract features (i.e., nouns and noun phrases) from the abstracts of the references cited by the user's documents of interest. The feature extraction phase is composed of literature retrieval and preprocessing steps. In the literature retrieval step, title of each reference will be parsed and used to retrieve their respective abstract from the literature database. In the preprocessing step, we use the rule-based part of speech tagger proposed by Brill (1994) to tag each word in the target abstract corpus and follow Voutilainen (1993) to develop a noun phrase parser for extracting nouns and noun phrases from each syntactically tagged abstract. Subsequently, the extracted nouns or noun phrases that belong to stopwords will be removed; otherwise, it will be stemmed to its base form.

## *Topic Discovery*

The topic discovery phase is to discover the latent topics existing in the abstracts of all references and the representative features (or terms) in each topic. The cited references are generally highly relevant to the citing paper and the topics among them can be diverse. To discover the topics inherent in the references could help identify the subjects or fields associated with the focal user's documents of interest, by which we might be able to accordingly assess and recommend subject-relevant documents for the focal user. In this study, we adopt LDA to discover the topics inherent in the references. The LDA is a generative probabilistic model for collections of discrete data such as text corpora. The basic assumption of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the topic distribution being a Dirichlet prior. A topic has probabilities of generating various words used to classify and interpret the topic, while a word may appear in several topics with different probability in each topic (Blei et al., 2003; Blei, 2012). Specifically, given a set of $i$ documents $D$, LDA discovers from them a set of $j$ topics $T$, and each of which consisted of a set of $k$ relevant words with its probability $w_{jk}$. In addition, LDA also derives the probability $\theta_{ij}$ of topic $t_j$ in documents $d_i$. For exampel, there are three documents $d_1$, $d_2$, and $d_3$. Table 1 shows the resultant topics discovered by LDA in them and the probability of each topic in document $d_1$. In the study, we derived the overall probability for a specific topic by average its

probability attained across all documents. That's, the overall probability of topic $t_1$ is calculated as $\varphi 1 = \frac{1}{||D||} \sum_{i=1}^{||D||} (\theta_i)$.

**Table1. Example of Probabilities of Discovered Topics Relevant to Document Set D**

| Topic | $\theta$ | Words in Topic | $w$ |
|---|---|---|---|
| $t_1$ | 0.407 | recommend | 0.10 |
| | | collaboration | 0.20 |
| | | filter | 0.05 |
| $t_2$ | 0.103 | ontology | 0.06 |
| | | approach | 0.04 |
| | | survey | 0.05 |
| $t_3$ | 0.184 | ontology | 0.07 |
| | | target | 0.30 |
| | | system | 0.22 |

### *Recommendation*

In recommendation phase, RTDR generates a list of candidate documents that may be relevant to the focal user's documents of preference. At previous phase, our proposed RTDR technique discovered reference topics and represented user's preference as a set of weighted topics where each topic comprises a set of weighted topic-related words (or features). We also represented each document in the literature database into the set of discovered topics and assigned the weight of each topic in a specific document by the cosine similarity between the probabilities of words in the topic and their frequencies in the document. Finally, RTDR determines the score of each document in the literature database by analyzing the similarity between the discovered reference topics and the topics in it. For example, assume that the term frequency of word "recommend", "ontology", and "filter" in document $d$ is 1, 2, and 4, respectively. The weight of topic $t_1$ in document $d$ is calculated as $\frac{1*0.1+0+4*0.05}{\sqrt{12+0+42*}\ \sqrt{0.12+0.22+0.052}} = 0.318$ and that of topic $t_2$ and topic $t_3$ are 0.684 and 0.185. While the weight of each topic in a document is calculated, the RTDR then determines its recommendation score by measuring the cosine similarity between discovered reference topics and topics in the document. The score is calculated as $\frac{0.407*0.318+0.103*0.684+0.184*0.185}{\sqrt{0.4072+0.1032+0.1842*}\ \sqrt{0.3182+0.6842+0.1852}} = 0.657$. Finally, the documents with the top-$k$ highest score will be recommended.

## Experimental Evaluation

### *Data collection*

This study recruited 30 graduate students who are major in information systems and working on their master's thesis. Each of the participants was asked to select 25 important research articles from the reference list of his or her own thesis and label five of them as the most important references. We tried to collect the title and abstract of these research articles and that of their respective references. Because of the overlap and the unavailability, we finally collected a total of 741 research articles and 2343 references cited by the collected articles.

### *Performance Measure*

Four evaluation metrics, including precision, recall, reciprocal rank (RR) (Deshpande & Karypis, 2004), and average precision (AP) (Chowdhury, 2010) are adopted to evaluate the performance of the investigated recommendation techniques. Precision and recall is defined as $\frac{|RA|}{|TR|}$ and $\frac{|RA|}{|TA|}$ where $|RA|$ is the number of recommended articles that are in the gold standard, $|TR|$ is the total number of articles that the investigated technique recommends, and $|TA|$ is the total number of articles in the gold standard. Because precision and recall rates are set-based metrics, we further adopted reciprocal rank (RR) and average precision (AP) to examine the effectiveness of the investigated techniques in

recommendation ranking. The RR is similar to uninterpolated precision (Deshpande & Karypis, 2004) and is defined as the sum of the reciprocal rank of all correct answers. For a query $q$, the value of reciprocal rank is $1/i_q$, where $i_q$ is the position of the relevant results for $q$; the value of reciprocal rank is zero if no relevant result exists. The AP metric takes into consideration the ranking and the position of the recommended articles and is defined as $\frac{\sum_{r \in R} P(r)}{|R|}$ where, $R$ is the order list of recommended articles that are in the gold standard, $|R|$ is the total number of articles in $R$, and $P(r)$ is the order of $r$ in the list over its recommended ranking. For example, assume 20 articles are recommended; among which, four articles a, b, c, d match user's preference and their ranking is 1, 2, 4, 10 in the list of recommendations. We thus can get $P(a)=1/1$, $P(b)=2/2$, $P(c)=3/4$, $P(d)=4/10$, and $AP=(1+1+0.75+0.4)/4=0.7875$.

### *Experiment design*

We design an experiment to evaluate the effectiveness of our proposed RTDR recommendation technique. The five most important research articles (with their references) chosen by each participant are taken as his/her preference profile (i.e., the training set of documents) and the remaining 20 research articles are viewed as the gold standard; i.e., the articles that RTDR technique have to identify from the collective research articles and make appropriate recommendations. Because our study intends to examine the effects of reference information, and thus, we adopt as the performance benchmark the traditional content-based recommender system (CBRS), which make document recommendations by analyzing the similarity of representative features (terms) between the user preference profile and the document to be recommended. Finally, the overall performance of each technique is calculated by the average performance across the 30 subjects.

### *Parameter Setting*

This study uses JGibbLDA, an open source software, to conduct LDA analysis (Phan & Nguyen, 2006). JGibbLDA adopts Gibbs sampling estimation to develop the topic model of the selected documents. A number of parameters need to be set in JGibbLDA, including the number of topics, the number of terms in a topic, hyper-parameters $\alpha$ and $\beta$. The output of JGibbLDA is a text file, which contains word-topic and topic-document distributions (Blei et al., 2003; Griffiths & Steyvers, 2004; Griffiths & Steyvers, 2004). The parameter setting of JGibbLDA we adopted in this study is shown in Table 2.

**Table 2. Parameter Setting for JGibbLDA**

| Parameter | Format | Description | Value |
|---|---|---|---|
| -alpha | double | LDA hyper-parameter alpha | 10 |
| -beta | double | LDA hyper-parameter beta | 0.1 (default) |
| -ntopics | int | Num. of topics | 5 |
| -savestep | int | The step at which the LDA model is saved to hard disk | 200 (default) |
| -twords | int | Num. of most likely words for each topic | 20 |
| -niters | int | Num. of Gibbs sampling iterations to continue estimating | 2,000 (default) |

## Evaluation Results

As shown in Table 3, our proposed RTDR outperforms CBRS across all performance metrics. However, the precision and the recall achieved by both techniques don't seem to arrive at a satisfying level. The RTDR attained significantly higher scores thank traditional CBRS technique in the two ranking-relevant metrics; i.e., reciprocal rank and average precision. The results of RR and AP suggest that the RTDR usually can get the first hit after the third recommendation; while CBRS gets the first hit after the sixth recommendation. Overall, the performance of the RTDR technique is advantageous over that of traditional CBRS technique. The RTDR can make better recommendations and recommend more documents that fit researchers' needs.

**Table 3. Comparative Evaluation Results**

| | CBRS | RTDR |
|---|---|---|
| Precision | 0.097 | 0.147 |
| Recall | 0.099 | 0.151 |
| Reciprocal Rank | 0.167 | 0.382 |
| Average Precision | 0.086 | 0.238 |

# Conclusion

This study proposed a reference topic-based document recommendation technique based on the user's preference documents and the reference information in them. Our proposed RTDR technique adopts LDA approach to discover the topics from the references and make recommendations on the basis of the reference topics. The experimental results indicated that RTDR outperformed the traditional CBRS technique. The topics discovered by LDA approach from titles and abstracts of the academic articles and their references can improve the performance of the recommender system.

There are few limitations that may influence the overall generalizability of this study. In this experiment, only journal articles are collected and used. In addition to the journal articles, articles from conferences, book chapters, newspapers, or website materials shall also be considered. As a result, the performance of the recommender system can be improved and its range can be expanded if these articles are complete gathered. In addition, this study aims to construct a recommender system for academic research articles; therefore, the participants recruited are who familiar with the academic research. There will have various applications if participants from distinct domains are recruited and there will have more interesting potential topics.

# 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," *IEEE Transactions on Knowledge and Data Engineering* (17:6), pp. 734-749.

Alspector, J., Kolcz, A., and Karunanithi, N. 1998. "Comparing Feature-Based and Clique-Based User Models for Movie Selection," in *Proceedings of the Third ACM Conference on Digital Libraries*, I. Witten, R. Akscyn, F.M. Shipman (Eds.), NewYork: ACM Press, pp. 11–18.

Ampazis, N. 2008. "Collaborative Filtering via Concept Decomposition on the Netflix Dataset," in *Proceedings of the 18th European Conference on Artificial Intelligence: Workshop on Recommender Systems (ECAI 2008)*, M. Ghallab, C.D. Spyropoulos, N. Fakotakis, N. Avouris (Eds.), Amsterdam: IOS Press, pp. 26–30.

Balabanovic, M. and Shohan, Y. 1997. "Fab: Content-Based, Collaborative Recommendation," *Communications of the ACM* (40:3), pp. 66-72.

Basilico, J. and Hofmann, T. 2004. "Unifying Collaborative and CBF," in *Proceedings of the 21st International Conference on Machine learning*, C. Brodley (Ed.), New York: ACM Press, pp. 65–72.

Beyah, G., Xu, P., Woo, H.G., Mohan, K., and Straub, D. 2003. "Development of An Instrument to Study the Use of Recommendation Systems," in *Proceedings of the Ninth Americas Conference on Information Systems*, A. Hevner, P. Cheney (Eds), Association of Information Systems, pp. 269–279.

Billsus, D. and Pazzani, M. J. 1998. "Learning Collaborative Information Filters," in *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, WI: Morgan Kaufmann Publishers Inc., pp. 46-54.

Blei, D.M. 2012. "Probabilistic Topic Models," *Communication of the ACM* (55:4), pp. 77-84.

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. "Latent Dirichlet allocation," *Journal of Machine Learning Research* (3:4-5), pp. 993–1022.

Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. 1999. "Partitioning-based Clustering for Web Document Categorization," *Decision Support Systems* (27:3), pp. 329-341.

Breese, J.S., Heckerman, D., and Kadie, C. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, Madison, WI, pp. 43-52.

Brill, E. 1994. "Some Advances in Rule-based Part of Speech Tagging," in *Proceedings of the Twelfth*

*National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, AAAI Press, pp. 722-727.
Burke, R. 2002. "Hybrid Recommender Systems: Survey and Experiments," *User Modeling and User-Adapted Interaction* (12:4), pp. 331–370.
Cao, Y. and Li, Y. 2007. "An Intelligent Fuzzy-Based Recommendation System for Consumer Electronic Products," *Expert Systems with Applications* (33:1), pp. 230-240.
Chen, P.Y., Wu, S.Y., and Yoon, J. 2004. "The Impact of Online Recommendations and Consumer Feedback on Sales," in *Proceedings of International Conference on Information System*, Washington, D. C., pp. 711–723.
Cheng, T.H. and Hu, P. 2007. "Content-based Recommendations Using Positive-Only Examples: A Single-Class Learning Approach," in *Proceedings of Sixth Workshop on e-Business*, H. Zhang, K.R. Lang, M. Parameswaran (Eds.), Tucson: Westing Publishing Inc., pp. 857–868.
Choi, S.H. and Ahn, B.S. 2011. "Rank Order-Based Recommendation Approach for Multiple Featured Products," *Expert Systems with Applications* (38:6), pp. 7081-7087.
Chowdhury, G. 2010. *Introduction to modern information retrieval*, Facet publishing.
Cronin, B. 1984. *The Citation Process: the Role and Significance of Citations in Scientific Communication*, London: Taylor Graham.
Culnan, M.J. 1986. "The Intellectual Development of Management Information Systems, 1972–1982: A Co-citation Analysis," *Management Science* (32:2), pp. 156–172.
Cutting, D., Karger, D., Pedersen, J., and Tukey, J. 1992. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," in *Proceedings of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318-329.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.A. 1990. "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* (41:6), pp.391-407.
Deshpande, M. and Karypis, G. 2004. "Item-based top-n recommendation algorithms," *ACM Transactions on Information Systems* (22:1), pp. 143-177.
Dumais, S., Platt, J., Heckerman, D., and Sahami, M. 1998. "Inductive Learning Algorithms and Representation for Text Categorization," in *Proceedings of the ACM 7th International Conference on Information and Knowledge Management*, Bethesda, MD, pp. 148-155.
Fogarolli, A. and Ronchetti, M. 2008. "Extracting Semantics from Multimedia Content," *Scalable Computing: Practice and Experience* (9), pp. 259–269.
Goldberg, D., Nichols, D., Oki, B., and Terry, D. 1992. "Using Collaborative Filtering to Weave An Information Tapestry," *Communications of the ACM* (35:12), pp. 61–70.
Guerrero Bote, V. P., de Moya Anegón, F., and Herrero Solana, V. 2002. "Document Organization Using Kohonen's Algorithm," *Information Processing and Management* (38:1), pp. 79-89.
Herlocker, J. and Konston, J. 2001. "Content-Independent Task-Focused Recommendation," *IEEE Internet Computing* (5:6), pp. 40–47.
Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. 1999. "An Algorithmic Framework for Performing Collaborative Filtering," in *Proceedings of the 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, F. Gey, M. Hearst, R. Tong (Eds.), New York: ACM Press, pp. 230–237.
Huang, M.H. 1997. "The Development of Relevance in Information Retrieval (in Chinese)," *Journal of Library and Information Studies* (12), pp. 39-62.
Hwang, S.Y., and Chuang, S.M. 2004. "Combining Article Content and Web Usage for Literature Recommendation in Digital Libraries," *Online Information Review* (28:4), pp. 260–272.
Hwang, S.Y., Wei, C.P., and Liao, Y.F. 2010. "Coauthorship Networks and Academic Literature Recommendation," *Electronic Commerce Research and Applications* (9:4), pp. 323-334.
Jinha, A.E. 2010. "Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence," *Learned Publishing* (23:3), pp. 258-263.
Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J. 1997. "Applying collaborative filtering to Usenet news," *Communications of the ACM* (40:3), pp. 77–87.
Krulwich, B. and Burkey, C. 1996. "Learning User Information Interests through Extraction of Semantically Significant Phrases," in *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, M. Hearst, H. Hirsh (Eds.), Menlo Park: AAAI Press, pp. 110–112.
Ku, Y.C., Lee, Y.H., and Lin, C.Y. 2012. "Use of Implicit User Feedbacks to Support Semantic-based Personalized Document Recommendation," in *Proceedings of the 11th Workshop on e-Business (WEB2012)*, Orlando, Florida.
Kuhlthau, C. 1993. Seeking Meaning: A Process Approach to Library and Information Services, Norwood, NJ: Ablex Publishing Co.
Lang, K. 1995. "NewsWeeder: Learning to Filter Netnews," in *Proceedings of the 12th International*

*Conference on Machine Learning*, A. Prieditis, S.J. Russell (Eds.), San Francisco: Morgan Kaufmann, pp. 331–339.

Larsen, B. and Aone, C. 1999. "Fast and Effective Text Mining Using Linear-time Document Clustering," in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16-22.

Lee, B.K. and Lee, W.N. 2004. "The Effect of Information Overload on Consumer Choice Quality in An On-Line Environment," *Psychology & Marketing* (21:3), pp. 159-183.

Lee, Y.-H., Yang, C.-S., Liau G.-Y. 2012. "A Social-Tag-Based Query Expansion Approach for Supporting Video Retrieval in Video Sharing Websites," *Journal of Information Management* (19:3), pp. 533-565.

Liang, T.P., Lai, H.J., Ku, Y.C. 2007. "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs," *Journal of Management Information Systems* (23:3), pp. 45–70.

Liang, T.P., Yang, Y.F., Chen, D.N., and Ku, Y.C. 2008. "A Semantic-Expansion Approach to Personalized Knowledge Recommendation," *Decision Support Systems* (45:3), pp. 401–412.

Linden, G., Smith, B., and York, J. 2003. "Amazon.com Recommendations: Item-to-item Collaborative Filtering," *IEEE Internet Computing* (7:1), pp. 76–80.

Liu, Y., Huang, X., and An, A. 2007. "Personalized Recommendation with Adaptive Mixture of Markov Models," *Journal of the American Society for Information Science and Technology* (58:12), pp. 1851–1870.

Martyn, J. 1964. "Bibliographic coupling," *Journal of Documentation* (20:4), pp. 236.

Middleton, S.E., de Roure, D., and Shadbolt, N.R. 2009. "Ontology-Based Recommender Systems," in *Handbook on Ontologies, 2nd edition*, S. Staab, R. Studer (Eds.), Berlin, Heidelberg: Springer-Verlag, pp. 779–796.

Mobasher, B., Dai, H., Luo, T., Sung, Y., and Zhu, J. 2000. "Integrating Web Usage and Content Mining for More Effective Personalization," in *Proceedings of the International Conference on E-Commerce and Web Technologies*, London, UK, pp. 165–176.

Mooney, R.J. and Roy, L. 2000. "Content-Based Book Recommending Using Learning for Text Categorization," in *Proceedings of the Fifth ACM Conference on Digital Libraries*, P. Nurnberg, D.L. Hicks, R. Furuta (Eds.), New York: ACM Press, pp. 195–204.

Mullner, R. and Chung, K. 2006. "Current Issues in Health Care Informatics," *Journal of Medical Systems* (30:1), pp. 1-2.

Pazzani, M. and Billsus, D. 1997. "Learning and Revising User Profiles: the Identification of Interesting Web Sites," *Machine Learning* (27:3), pp. 313–331.

Pazzani, M. 1999. "A Framework for Collaborative, Content-Based and Demographic Filtering," *Artificial Intelligence Review* (13:5-6), pp. 393-408.

Phelan, O., McCarthy, K., and Smyth, B. 2009. "Using Twitter to Recommend Real-Time Topical News," in *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*, R. Burke, A. Felfernig, L. Schmidt-Thieme (Eds.), New York: ACM Press, pp. 385–388.

Polat, H. and Du, W. 2008. "Privacy-Preserving Top-N Recommendation on Distributed Data," *Journal of the American Society for Information Science and Technology* (59:7), pp. 1093–1108.

Resnick, P. and Varian, R.H. 1997. "Recommender Systems," *Communications of the ACM* (40:3), pp. 56–58.

Roussinov, D. and Chen, H. 1999. "Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques," *Decision Support Systems* (27:1), pp. 67-79.

Schafer, J.B., Konstan, J., and Riedl J. 1999. "Recommender System in E-commerce," in*Proceedings of the first ACM conference on electronic commerce*, S. Feldman, M. Wellman (Eds), New York: ACM Press, pp. 158–166.

Schutze, H. and Silverstein, C. 1997. "Projections for Efficient Document Clustering," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, pp. 74-81.

Semeraro, G., Basile, P., de Gemmis, M., and Lops, P. 2007. "Content-based Recommendation Services for Personalized Digital Libraries, Digital Libraries: Research and Development," *Lecture Notes in Computer Science* (4877), pp. 77–86.

Senecal, S. and Nantel, J. 2004. "The Influence of Online Product Recommendations on Consumers' Online Choices," *Journal of Retailing* (80:2), pp. 159–169.

Shardanand, U. and Maes, P. 1995. "Social Information Filtering: Algorithms for Automating Word of Mouth," in *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, I.R. Katz, R. Mack, L. Marks, M.B. Rosson, J. Nielsen (Eds.), New York: ACM Press, pp. 210–217.

Spangler, S., Kreulen, J.T., and Lessler, J. 2003. "Generating and Browsing Multiple Taxonomies Over A Document Collection," *Journal of Management Information Systems* (19:4), pp. 191-212.

Srivastava, J., Cooley, R., Deshpande, M., and Tang, P. 2000. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations* (1:2), pp. 12–23.

Steyvers, M. and Griffiths, T. 2006. "Probabilistic topic models. Latent Semantic Analysis: A Road to Meaning," in Lawrence Erlbaum, T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds).

Tang, C., Dwarkadas, S., and Xu, Z. 2004. "On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, South Yorkshire, UK, pp.112-121.

Vakkari, P. 2003. "Changes of Search Terms and Tactics While Writing a Research Proposal: A Longitudinal Ccase Study," *Information Processing and Management* (39:3), pp. 445-463.

Voutilainen, A. 1993. "NPtool: A Detector of English Noun Phrases," in *Proceedings of Workshop on Very Large Corpora*.

Wang, P. and Sogergel, D. 1998. "A Cognitive Model of Document Use during A Research Project. Study I. Document Selection", *Journal of the American Society for Information Science and Technology* (49:2), pp. 115-133.

Wei, C., Yang, C.S., Hsiao, H.W., and Cheng, T.H. 2006. "Combining Preference- and Content-based Approaches for Improving Document Clustering Effectiveness," *Information Processing and Management* (42:2), pp. 350-372.

Wei, C.P., Chen, Y., Yang, C., and Yang, C.C. 2010. "Understanding What Consumers Concern: A Semantic Approach for Product Feature Extraction from Consumer Reviews," *Journal of Information Systems and E-Business* (8:2), pp. 149–167.

Wei, C.P., Shaw, M.J., and Easley, R.F. 2002. "A Survey of Recommendation Systems in Electronic Commerce," in *E-Service: New Directions in Theory and Practice*, R.T. Rust, P.K. Kannan (Eds.), Armonk: M.E. Sharpe Publisher, pp. 168–199.

Wilson, P. 1973. "Situational Relevance," *Information Storage & Retrieval* (9:8), pp. 457-471.

Xiao, B. and Benbasat, I. 2007. "E-commerce Product Pecommendation Agents: Use, Characteristics, and Impact," *MIS Quarterly* (31:1), pp. 137–209.

Yang, C.C., Tang, X., Wong, Y.C., and Wei, C.P. 2010. "Understanding Online Consumer Review Opinions with Sentiment Analysis Using Machine Learning," *Pacific Asia Journal of the Association for Information Systems* (2:3), pp. 73–89.

Zhang, X., Asano, Y., and Yoshikawa, M. 2010. "Analysis of Implicit Relations on Wikipedia: Measuring Strength through Mining Elucidatory Objects," in *DASFAA 2010*. LNCS, 5981, H. Kitagawa, Y. Ishikawa, Q. Li, C. Watanabe (Eds.), Heidelberg: Springer, pp. 460–475.

Zhen, L., Huang, G.Q., and Jiang, Z. 2010. "An Inner-Enterprise Knowledge Recommender System," *Expert Systems with Applications* (37:2), pp. 1703-1712.