Diversifying Search Results through Pattern-Based Subtopic Modeling

Wei Zheng, Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Hui Fang, Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Hong Cheng, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Xuanhui Wang, Facebook, Menlo Park, CA, USA

ABSTRACT

Traditional information retrieval models do not necessarily provide users with optimal search experience because the top ranked documents may contain excessively redundant information. Therefore, satisfying search results should be not only relevant to the query but also diversified to cover different subtopics of the query. In this paper, the authors propose a novel pattern-based framework to diversify search results, where each pattern is a set of semantically related terms covering the same subtopic. They first apply a maximal frequent pattern mining algorithm to extract the patterns from retrieval results of the query. The authors then propose to model a subtopic with either a single pattern or a group of similar patterns. A profile-based clustering method is adapted to group similar patterns based on their context information. The search results are then diversified using the extracted subtopics. Experimental results show that the proposed pattern-based methods are effective to diversify the search results.

Keywords: Clustering, Diversity, Frequent Pattern Mining, Information Retrieval, Subtopics

INTRODUCTION

It has been a long standing challenge to develop effective retrieval models that can provide users with optimal search experiences. Traditional retrieval models rank documents based on only their relevance scores and ignore the semantic relations among returned documents. However, different documents may contain the same piece of relevant information, and returning all of them would not be a good ranking strategy. Intuitively, search results covering different pieces of relevant information, i.e., query subtopics, are more desirable than those covering the same

DOI: 10.4018/jswis.2012100103

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

single piece of relevant information multiple times. Thus, it is necessary to rank documents based on not only relevance but also diversity.

Diversifying search results can benefit both queries with extrinsic diversity such as ambiguous queries and those with intrinsic diversity such as under-specified queries (Clarke, Craswell, & Soboroff, 2009; Craswell, Fetterly, Najork, Robertson, & Yilmaz, 2009; Radlinski, Bennett, Carterette, & Joachims, 2009). The general goal of result diversification is to return a list of relevant documents that cover all of the subtopics of a query. As an example, query "java" is ambiguous. Since this query could have several interpretations and we may not know which interpretation reflects a user's need, it would be important to return diversified results covering different interpretations so that they can satisfy all possible information needs. Another example is that a user doing a literature survey uses query "computer architecture" to find documents that cover representative topics in computer architecture. The user would prefer a ranking of documents covering different topics in computer architecture while avoiding excessive redundancy. The state of the art diversification methods aim to diversify search results so that they can cover more query subtopics. Thus, one of the key challenges is how to identify semantically meaningful subtopics for a given query.

In this paper, we propose to use the frequent pattern mining approach to identify query subtopics from a document set. We design a novel result diversification framework that models the diversity explicitly through pattern-based subtopics. The basic idea is to combine the relevance, through existing relevance-based retrieval models, with the diversity, through pattern-based subtopic modeling. In particular, we define a pattern as a set of semantically related and meaningful terms extracted from documents. For example, a pattern could be a phrase or a term collocation such as "programming language", "code developer" and "gourmet coffee". Such patterns can be efficiently discovered with the state-of-the-art frequent pattern mining algorithms (Bayardo, 1998; Han, Pei, & Yin, 2000; Zaki, 2000). We propose to model a query subtopic by *a single pattern* that is relevant to the query. However, since different patterns could be semantically related and complementary, they can be merged to form a more complete semantic unit. Thus, as an alternative, we model a query subtopic as *a group of semantically related patterns* that are relevant to the query. For example, by grouping the two patterns "programming language" and "code developer", the subtopics of query "java" could be "programming language code developer" and "gourmet coffee". To discover the related pattern groups as query subtopics, we use a profile-based pattern clustering method to group semantically related patterns (Yan, Cheng, Han, & Xin, 2005). The method was originally designed to summarize frequent itemsets into different groups so that similar item sets are assigned to the same group. The similarity between two patterns is measured based on not only the pattern composition, i.e., terms contained in the patterns, but also the context of patterns, i.e., documents containing the patterns. In this work, we represent the context of a pattern with a profile, which is the term distribution of the document set that contains the pattern. The similarity between two patterns can then be measured by the divergence between their profiles and similar patterns are grouped together to model one query subtopic.

Our main contribution is the novel methods to model query subtopic based on a frequent pattern mining approach, which effectively identifies query subtopics based on term co-occurrences. The pattern-based subtopic modeling methods allow us to focus on the important content of the documents and are more robust to the noises in the documents. To the best of our knowledge, no existing work on search result diversification used the pattern-based subtopic modeling idea. Furthermore, it provides a demonstration of how data mining techniques, in particular, frequent pattern mining, can help solve information retrieval problems.

The rest of the paper is organized as follows. We first discuss related work and present an overview of the proposed pattern-based

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

methods for result diversification. We then explain how to extract patterns from a document collection and how to model query subtopics using the extracted patterns. We also explain the implementation details and analyze the time complexities of proposed methods. Finally, we discuss experiment results and conclude.

RELATED WORK

The earliest study of result diversification can be traced back to the early sixties (Goffman, 1964). Since then, many studies have tried to rank documents based on not only relevance but also diversity (Agrawal, Gollapudi, Halverson, & Ieong, 2009; Boyce, 1982; Carbonell & Goldstein, 1998; Carterette & Chandar, 2009; Chen & Karger, 2006; Gollapudi & Sharma, 2009; Radlinski et al., 2009; Radlinski & Dumais, 2006; Yue & Joachims, 2008; Zhai, Cohen, & Lafferty, 2003; Zheng, Xuanhui, Fang, & Cheng, 2012).

The proposed methods can be classified into two categories based on how the diversity is modeled. The first category is the redundancy-based method, where the diversity is modeled through the relations among documents and the goal is to minimize the redundant information among the retrieved documents (Carbonell & Goldstein, 1998; Chen & Karger, 2006; Gollapudi & Sharma, 2009; Yue & Joachims, 2008; Zhai et al., 2003). For example, Carbonell and Goldstein (1998) proposed the maximal marginal relevance (*MMR*) ranking strategy to balance the relevance and the redundancy. Motivated by this work, Zhai et al. (2003) used statistic language models to balance the relevance and redundancy. Chen and Karger (2006) presented a sequential document selection algorithm to optimize an objective function that aims to find at least one relevant document for every user. Yue and Joachims (2008) studied a learning to rank algorithm to retrieve relevant documents covering maximally distinct words. The second category is the subtopic-based method (Agrawal et al., 2009; Radlinski & Dumais, 2006; Carterette & Chandar, 2009). The goal is to maximize the coverage of query subtopics in the retrieved documents where the subtopic identification is an important step. Radlinski and Dumais (2006) used the reformulated queries from a query log as subtopics of a query. Agrawal et al. (2009) classified queries and documents into different subtopics according to existing taxonomies. Carterette and Chandar (2009) discovered query subtopics using either topic modeling such as LDA (Blei, Ng, & Jordan, 2003) or relevance modeling (Lavrenko & Croft, 2001).

The launch of the diversity task in TREC 2009 Web track has established a common test bed for result diversification and attracted a lot of attention in the research community (Clarke et al., 2009). Some participants used redundancy-based methods and removed redundant information based on either the document content (Balog et al., 2009) or the host information of the documents (Craswell et al., 2009). Others used subtopic-based methods and discovered query subtopics using different resources such as the document content (Dou et al., 2009; Li et al., 2009; Bi et al., 2009; Balog et al., 2009), anchor text (Dou et al., 2009), host websites (Dou et al., 2009) and query suggestions from Web search engines (Li et al., 2009; Balog et al., 2009; Mccreadie, Macdonald, Ounis, Peng, & Santos, 2009).

Compared with the previous work, we propose a novel subtopic-based method. Our work differs from the previous work in that: (1) we attempt to model query subtopics with salient patterns extracted from relevant documents; and (2) we apply an efficient maximal frequent pattern mining algorithm proposed by Bayardo (1998) to discover patterns and use a profile-based approach proposed by Yan et al. (2005) to cluster the patterns into groups for modeling different query subtopics. Frequent pattern mining has been an important research topic in data mining community for over a decade, and many efficient algorithms have been proposed (Agrawal, Imieliński, & Swami, 1993; Agrawal & Srikant, 1994; Han et al., 2000; Zaki, 2000; Bayardo, 1998). Intuitively, these algorithms should be useful to discover interesting patterns,

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

i.e., semantically meaningful text units, from document collections. However, as far as we know, our work is the first study trying to apply the pattern mining algorithms to search result diversification. Our work demonstrates how information retrieval problems can benefit from data mining techniques, and introduces a new application scenario of frequent pattern mining.

AN OVERVIEW OF PATTERN-BASED RESULT DIVERSIFICATION

The goal of result diversification is to return a list of relevant documents that cover all of the subtopics of a query while avoiding excessive redundancy in the top ranked results (Clarke et al., 2009). For example, given query "java", a system should return relevant documents about not only programming language but also coffee and island, since such results with mixed subtopics could provide users with a more complete picture of the relevant information. Thus, one of the key problems is to discover subtopics of a query from a document collection.

Most related work on subtopic-based result diversification relies on document clustering to discover subtopics of the query (Clarke et al., 2009; Dou et al., 2009; Bi et al., 2009; Li et al., 2009; Balog et al., 2009). However, a document may contain both relevant and non-relevant information. When the non-relevant information is long in the document, it may significantly affect the document clustering results, which leads to the incorrect identification of clusters, i.e., subtopics. To overcome this limitation, we propose a pattern-based search result diversification framework which directly models query subtopics with important patterns extracted from retrieved documents, and diversifies the documents according to the pattern-based subtopics. We now discuss two main challenges and briefly explain how to address each of them.

The first challenge is how to define and extract patterns that are related to the subtopics of a query. A pattern, in general, could be any semantic features extracted from documents such as a set of terms or a sequence of terms. We propose to define a pattern as a semantically meaningful text unit extracted from relevant documents for a given query. Intuitively, every pattern is a group of semantically related terms that can represent part of the relevant information. For example, for query "java", the patterns extracted from relevant documents might be "programming language", "code development" and "coffee flavor". Since co-occurrences of terms usually indicate that there exist semantic relationships between the terms (Berger & Lafferty, 1999; Schütze & Pedersen, 1997), we formally define a *pattern* as a set of terms that co-occur frequently in a relevant document collection, which could be either true or pseudo relevant documents such as top ranked search results. According to the formal definition of patterns, we propose to adapt a maximal frequent item set mining algorithm Max-Miner proposed by Bayardo (1998) to efficiently extract the patterns.

The second challenge is how to model and discover query subtopics using the extracted patterns. We propose two solutions. The first one is to model a query subtopic with a single pattern, which is referred to as *single pattern based method*. Specifically, we rank all the patterns based on their importance and select top ranked patterns as query subtopics. However, patterns could be related and a group of similar patterns may correspond to one subtopic of a query. For example, "program language" and "code development" are related and they correspond to the same subtopic of the query "java". Thus, the second solution is to model a subtopic for a query as a group of related patterns, which is referred to as *pattern cluster based method*. To group semantically related patterns into the same clusters, we propose to use a profile-based clustering method (Yan et al., 2005) that uses both the content and context information of a pattern to cluster the extracted patterns. Every pattern cluster then corresponds to a query subtopic.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

PATTERN EXTRACTION

A pattern, in general, can be defined as a semantically meaningful text unit extracted from documents. Since the co-occurrences of terms often indicate semantic relationships between the terms (Berger & Lafferty, 1999; Schütze & Pedersen, 1997; Fang & Zhai, 2006), we define a pattern as a set of terms that frequently co-occur in a document collection. However, it is very time consuming to compute the similarity between every term pair and cluster them. We therefore borrow the concept of maximal frequent itemset (Bayardo, 1998) in data mining and give a more rigid definition of patterns. We also discuss how to extract such patterns from a set of documents. The formal definition of patterns is given as follows.

DEFINITION(PATTERN) Let $D=\{d_1,d_2,...,d_n\}$ be a set of documents and $V=\{t_1,t_2,...,t_m\}$ be the vocabulary set. A set of terms s ($s \subseteq V$) is defined as a **pattern** if $|D_s| \geq$ min_supp and there exists no superset s' ($s' \subseteq V$) such that $s \subset s'$ and $|D_{s'}| \geq$ min_supp, where D_s is the set of documents in D that contain all terms in s, $|D_s|$ is the number of documents in D_s and min_supp is the user-specified threshold.

The definition suggests that a set of terms is a pattern when it satisfies the following two requirements: (1) the terms need to co-occur no less than *min_supp* times in the document set *D*; and (2) there exists no superset in which all of the terms co-occur no less than *min_supp* times in the document set. The first requirement, i.e., the minimum support threshold, ensures that a pattern contains a group of semantically related terms. The underlying assumption is that if a group of terms co-occur frequently in a document collection, they are semantically related. The second requirement, which essentially corresponds to the maximal itemset definition in data mining, allows us to focus on larger patterns, i.e., the ones with more terms, without being overwhelmed by the smaller ones with redundant information, because the requirement excludes all subsets of the maximal patterns from the output.

Table 1 shows an example document collection with 7 documents. The vocabulary set is the union of all terms appearing in Table 1. We assume that *min_supp* is set to 2. Since four terms {*time, magazine, family, tree*} co-occur in two documents D_1 and D_3, both the set of these four terms and all of its subsets satisfy the first requirement, i.e., *min_supp* threshold. Actually there are many term sets satisfying the first requirement in the collection. On the contrary, there are only four patterns, i.e., four term sets satisfying both requirements, in the collection: {*time, magazine, family, tree*}, {*photo, essay, family, tree, time, barack*}, {*biographical, mother, obama*} and {*shall, soon, obama, tree, good*}. It is clear that the second requirement of

Table 1. An example of the document collection

IDs	Documents
D_1	time, magazine, family, tree, article, newsweek, claim, …
D_2	photo, essay, family, tree, time, barack, post, state, …
D_3	photo, essay, family, tree, time, barack, magazine, …
D_4	biographical, mother, obama, grandmother, hawaii, …
D_5	biographical, mother, obama, father, genealogist, …
D_6	provide, good, obama, shall, soon, tree, …
D_7	good, purchase, obama, shall, soon, tree, …

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

maximal pattern in the definition allows us to focus on a small number of larger semantically related term groups.

We now discuss how to efficiently extract the defined patterns from a document collection. In fact, if we assume that a term is an item and a document is a transaction in databases, the definition of patterns essentially corresponds to that of the maximal frequent itemsets studied in the data mining community. Thus, we propose to adapt a maximal frequent itemset mining algorithm, i.e., Max-Miner proposed by Bayardo (2009), for pattern extraction. Specifically, given a document collection, we construct a set-enumeration tree over all the terms. Each node is a pattern candidate. We then perform a breadth-first search to find the patterns. Max-Miner uses two pruning strategies, i.e., superset frequency pruning and subset infrequency pruning, to delete nodes that are impossible to be patterns and reduce the search space. If a superset candidate is frequent, all its subset will be pruned from the tree. If a subset candidate is infrequent, all its superset will also be pruned from the tree. In particular, this algorithm scales roughly linearly in the number of patterns and the size of document collection.

Since the goal of this work is to model subtopics of a query with patterns, we focus on only *relevant patterns*, i.e., those extracted from a set of relevant documents of the query. The rationale is that terms related to a query subtopic usually co-occur in the relevant documents with a reasonable frequency. However, one document may cover more than one subtopic. It is difficult to separate different subtopics when we use the document as the unit to extract patterns. In order to solve this problem and to take advantage of the term proximity, we decide to use fixed-length segments instead of documents as transactions for pattern extraction. Moreover, following the ideas of pseudo feedback, we assume that top ranked segments are relevant and use them to construct the relevant transaction set.

In summary, there are three steps to extract relevant patterns from a document collection for a given query. First, we retrieve segments that are relevant to the query from the collection. Second, each segment can be represented as a transaction based on the words in the segment. We can then apply Max-Miner methods to extract maximal patterns from the top ranked segments.

SUBTOPIC MODELING

In this section, we describe how to model subtopics of a query based on the extracted patterns. Assuming that a subtopic of a query is a group of semantically related terms representing an aspect or an interpretation of the query, we explore two ways of modeling query subtopics. The first one is to assume that a query subtopic can be represented as a single relevant pattern, which is referred to as *single pattern based method*. The second one is to assume that a subtopic can be modeled as a group of relevant patterns that are related to each other, which is referred to as *pattern cluster based method*. Since the first method is straightforward, we now give more details for the second method, i.e., discover the subtopics by clustering the extracted patterns. We then discuss how to rank the subtopics based on their importance.

Pattern Cluster based Subtopic Discovery

We propose to cluster related patterns together to find better representations of query subtopics. Although a pattern is a group of semantically related terms, different patterns might also be related and the related patterns may form a more complete semantically meaningful unit. For example, in Table 1, we can find two patterns s={*time, magazine, family, tree*} and s'={*photo, essay, family, tree, time, barack*}. s is generated from the document set D_s={D_1, D_3} while s' is generated from $D_{s'}$={D_2, D_3}. As $D_s \neq D_{s'}$, s and s' are output as two patterns, rather than a merged one as $s \cup s'$={*time, magazine, family, tree, photo, essay, barack*}, which is a more complete semantically meaningful unit.

An important factor in clustering is the distance measure. Given two patterns, we could measure their distance only based on their over-

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

lapped terms. However, if these two patterns do not share many terms but are related to the same query subtopic, i.e., they complement each other, their distance would be very high and they may be incorrectly partitioned into two different clusters. Thus, a more reasonable distance should be computed based on not only the content of patterns, i.e., terms contained in the patterns, but also the context of patterns, i.e., documents containing the patterns.

In this work, we propose to apply a profile-based clustering approach (Yan et al., 2005) to cluster the patterns based on their context. We first build the profile of each pattern using the documents containing the pattern. The basic idea is to capture the context of a pattern s through a profile, i.e., the term distribution of the document set D_s that generates the pattern s. Let $V=\{t_1,...,t_m\}$ denote the vocabulary. Formally, we represent the context profile of a pattern s as a term probability distribution vector computed from D_s, i.e.:

$$CP(s) = (p(t_1), p(t_2), ..., p(t_m)),$$

where

$$p(t_k) = \frac{\sum_{d \in D_s} c(t_k, d)}{\sum_{d \in D_s} | d |}.$$

Note that $p(t_k)$ in fact is the maximum likelihood estimation (MLE) of term t_k in D_s, $c(t,d)$ is the occurrence of term t in document d and $|d|$ is the length of document d. Table 2 shows the context profile for the pattern $s=\{time, magazine, family, tree\}$ computed from $D_s=\{D_1, D_3\}$. Table 2 only shows the terms with non-zero probabilities.

Given the patterns and their profiles, we use *K-Means* clustering to group the set of patterns into K groups. In particular, K patterns are selected randomly as initial cluster centers. The remaining patterns are then assigned to one of the K clusters according to the distance measure. We measure the distance between two patterns s and s' based on the divergence between their context profiles $CP(s)$ and $CP(s')$. Specifically, we use the Kullback-Leibler divergence (Cover & Thomas, 1991) between the two context profiles as the distance function:

$$KL(CP(s) \,||\, CP(s')) = \sum_{k=1}^{m} p_s(t_k) \log \frac{p_s(t_k)}{p_{s'}(t_k)}$$

where $p_s(t_k)$ and $p_{s'}(t_k)$ are probabilities of term t_k in $CP(s)$ and $CP(s')$, respectively. To avoid zero probabilities, we smooth $p(t_k)$ with Jelinek-Mercer method (Zhai & Lafferty, 2001). This clustering process iterates until convergence, e.g., the cluster membership does not change much or there is small change in the cluster profiles.

After assigning the patterns into K clusters, we use each cluster as a subtopic of the query. Thus, every subtopic contains a set of semantically related patterns. We represent each subtopic with a profile of the corresponding cluster. Here a cluster profile is the term probability

Table 2. A context profile

Terms	Probability	Terms	Probability
Time	0.143	Newsweek	0.071
Magazine	0.143	Claim	0.071
Family	0.143	Photo	0.071
Tree	0.143	Essay	0.071
Article	0.071	Barack	0.071

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

distribution vector computed over the union of the supporting documents for every pattern in the cluster. The output is *K* subtopics with a list of terms and their probabilities. The value of *K* will be tuned in the experiment based on the diversification performance.

Subtopic Weighting

Pattern-based subtopics are essentially groups of semantically related terms. However, similar to terms, not every extracted pattern is equally important, and not every discovered query subtopic is equally important. For example, in Table 1, the discovered pattern {*shall, soon, obama, tree, good*} is less important than other patterns since it contains more common words.

The importance of a subtopic is important in the diversification process, because it is one of the factors that determine how to re-rank the retrieved documents.

Intuitively, the importance of a pattern, i.e., a query subtopic, is related to the weights of terms occurring in the pattern, i.e., the subtopic. Thus, we propose to compute the importance of a pattern or a query subtopic as follows:

$$weight(s) = \sum_{t \in s} weight(t)$$

where *t* denotes a term, *s* denotes a pattern or a query subtopic, and *weight(t)* is the weight of term *t*. We explore the following three term weighting strategies to compute *weight(t)*.

1. *Traditional IDF weighting* (Salton & Buckley, 1988; Zobel & Moffat, 1998): *IDF* assigns higher weights to terms that occur less frequently:

$$weight_{IDF}(t) = \log \frac{N}{df(t)} \quad (1)$$

2. *Term importance score* (Swaminathan, Mathew, & Kirovski, 2009): This metric assigns lower weights to terms with either high or low frequency. The rationale is that terms with high frequency could be common words and those with low frequency are not representative and could be misspelled terms:

$$weight_{Imp}(t) = \frac{df(t)}{N} \log \frac{N}{df(t)} \quad (2)$$

3. *Semantic similarity based weighting* (Fang & Zhai, 2006): This weighting strategy is to measure how closely related a term is to a query. It computes the weight of a term based on not only its semantic similarity with query terms but also the importance of the query terms:

$$weight_{sim}(t) = \frac{\sum_{t_q \in q} weight_{IDF}(t_q) \cdot sim(t_q, t)}{|q|} \quad (3)$$

Note that *q* is a query, $|q|$ is its length, t_q is a query term, *N* is the number of documents in the collection, *df(t)* is the number of documents containing term *t* and $sim(t_q,t)$ denotes the mutual information between terms (Fang & Zhai, 2006).

SYSTEM IMPLEMENTATION

To implement the diversification system, we follow a commonly used diversification strategy (Santos, Macdonald, & Ounis, 2010), which can be described as follows:

1. Given a query, the system first retrieves a list of relevant documents;
2. It identifies query subtopics from the retrieved documents;
3. The system then diversifies the results through re-ranking the documents based on their diversity scores.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

The first step can be implemented using any existing retrieval functions, and the third step can be implemented using any existing diversity functions. In this paper, we used Dirichlet Prior retrieval function (Zhai & Lafferty, 2001) to retrieve relevant documents in Step 1 and then use a state-of-the-art diversification function, i.e., *xQuAD* (Santos et al., 2010), in Step 3.

The basic idea of *xQuAD* is to iteratively select documents that are not only similar to the query but also similar to the important subtopics that have not been well covered by previously selected documents:

$$d^* = \arg\max_{d}((1-\lambda)P(d \mid q) + \lambda\sum_{s\in S} P(s \mid q)P(d \mid s)\prod_{d'\in D}(1-P(d' \mid s))) \quad (4)$$

where S is the set of subtopics, D is a set of documents that have been selected in the query, and λ is a parameter that controls the balance of the relevance and diversity scores. $P(s|q)$ denotes the importance of s given q. $P(d|q)$ and $P(d|s)$ measures the relevance scores of d with respect to q and s, and are computed using Dirichlet Prior retrieval function as well.

We now provide more details on how to perform the second step based on the proposed methods. First, we use the frequent itemset mining algorithm (Bayardo, 1998) to extract patterns from the retrieved segments. Second, we may assume that there are K subtopics in the query, and then apply one of the subtopic modeling methods to find the subtopics and compute their weights. Note that we assume that every query has a fixed number of subtopics in this work and leave the study of predicting the number of query subtopics as future work.

Specifically, for the *single pattern based method*, we assume that a query subtopic is modeled with a single pattern. We then rank all the extracted patterns using one of the weighting methods described in Equations (1-3) and select top K ranked patterns as the K subtopics. For the *pattern cluster based method*, we assume that a subtopic is modeled as a group of patterns. We use the profile-based clustering algorithm to group patterns into K clusters, each of which corresponds to a subtopic.

DISCUSSIONS ON TIME COMPLEXITY

Let us first consider the *single pattern based method*. It has two steps: extracting patterns and computing the weight for each pattern. Since the pattern extraction method scales roughly linearly in the number of maximal patterns irrespective of the length of the longest patterns (Bayardo, 1998), the time complexity of the pattern extraction step is $O(P \cdot R_S \cdot V_S)$, where P is the number of maximal patterns, R_S is the number of retrieved segments, and V_S is the vocabulary size of the segments, i.e., the number of different words in the segments. Since we have the index of segments, the time complexity of subtopic weighting using *IDF*, *term importance score* and *semantic similarity* is $O(P \cdot V_S)$, $O(P \cdot V_S)$ *and* $O(V_s \cdot Q \cdot R_S + P \cdot V_S)$, respectively, where Q is the number of terms in the query. Q is often smaller than P. Therefore, the overall time complexity of *single pattern based method* is $O(P \cdot V_S \cdot R_S)$.

We now consider the *pattern cluster based method*, which includes three steps: extracting patterns, clustering patterns and computing the weight for each cluster. Thus, the time complexity is $O(P \cdot V_S \cdot R_s + P \cdot V_S \cdot T \cdot K)$, where T is the number of iterations in the *K-means* clustering method and K is the number of query subtopics. R_S is larger than $T \cdot K$. Therefore, the time complexity is $O(P \cdot V_S \cdot R_S)$.

As discussed earlier, *PLSA*, i.e., Probabilistic Latent Semantic Analysis (Hofmann, 1999), is an existing method to extract query subtopics. The time complexity of subtopic extraction using *PLSA* is $O(T \cdot K \cdot V \cdot R_D)$ (Xue, Dai, Yang, & Yu, 2008), where R_D is the number of retrieved documents, V is the vocabulary size of documents.

Comparing the time complexity of the proposed methods with *PLSA*, we find that Vs is significantly smaller than V, Rs is similar to

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

R_D, and P, depending on the value of *min_supp,* can be smaller or larger than $T \cdot K$. Therefore, the time complexity of both *single pattern based method* and *pattern cluster based method* is similar with that of *PLSA*.

EXPERIMENTS

Experiment Design

We evaluate the effectiveness of the proposed pattern-based methods over the standard collections used for the diversity task in Web tracks of TREC09, TREC10 and TREC11 collections (Clarke et al., 2009; Clarke, Craswell, Soboroff, & Cormack, 2010; Clarke, Craswell, Soboroff, & Voorhees, 2011). There are 50 official topics in each collection, and we use only their query fields in our experiments. For the document collection, we use the ClueWeb09 "Category B" data set with 50 million English-language pages. The performance is measured using one of the official measures α *-nDCG* (α normalized Discounted Cumulative Gain) at three retrieval depths, i.e., top 5, top 10 and top 20 documents. We use α *-nDCG @20* as the primary evaluation measure. The preprocessing involves stemming with Porter's stemmer and stop word removal.

Note that we propose two methods to model subtopics with patterns, i.e., *SP* (single pattern-based subtopics) and *Cluster* (cluster-based subtopics), and three methods for subtopic weighting, i.e., *IDF* (traditional idf weighting), *Imp* (term importance score) and *Sim* (semantic similarity-based weighting). Thus, we totally have six pattern-based methods to be evaluated. We also implement three baseline methods. *No-Diversity* is the result ranking documents based on only relevance using the Dirichlet retrieval function (Zhai & Lafferty, 2001). *PLSA* denotes a state-of-the-art diversification method that uses *PLSA* to extract the subtopics and *xQuAD* to diversify the documents. *MMR* denotes a redundancy-based diversification method, i.e., Maximal Marginal Relevance (Carbonell & Goldstein, 1998), which re-ranks the documents based on the redundancy of a document with respect to the previously selected documents.

Effectiveness of Pattern-based Methods

We first conduct experiments to evaluate the effectiveness of the six proposed pattern-based diversity methods. Table 3 shows the optimal performance of all the compared methods on TREC collections, i.e., TREC09, TREC10 and TREC11. All the parameters in these methods are tuned in ranges and set to the optimum values that correspond to the optimal performance of the methods on each collection. The details of parameter tuning are explained in the next sub-section. We also report the significance test results based on Wilcoxon test (Mendenhall, Wackerly, & Schaeffer, 1990) at significance level 0.05 and 0.1.

In addition to the optimal performance, we also train the parameter values on one collection, i.e., TREC09, use the trained parameter values to diversity results on testing collections, i.e., TREC10 and TREC 11, and report the results in Table 4. We make the following four interesting observations.

First, most of the pattern-based diversity methods can consistently outperform the baseline methods. The test performance of *Cluster* methods in Table 4 can be ranked 2nd among the official diversity runs of TREC11 on Category B collection (Clarke et al., 2011). The method *Cluster+Sim* can significantly outperform baselines on TREC2010 and TREC2011 collections. The reason that the pattern-based methods outperform the baselines is that the pattern-based methods are less sensitive to the non-relevant documents in the original retrieval results than the baselines, i.e., *MMR* and *PLSA*. The non-relevant documents in the retrieval results are often less similar to relevant documents and the non-relevant documents themselves are also different (Zhai et al., 2003; Zheng & Fang, 2011). The pattern-based methods can focus on the most important information

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Table 3. Optimal performances of diversity methods. ▲, ◆ and ‡ denote the results are significantly better than results of NoDiversity, MMR and PLSA at 0.05 level in Wilcoxon test, respectively. △, ◇ and † denote the results are significantly better than results of NoDiversity, MMR and PLSA at 0.1 level in Wilcoxon test, respectively.

Methods		α -nDCG (TREC09)			α -nDCG (TREC10)			α -nDCG (TREC11)		
		@5	@10	@20	@5	@10	@20	@5	@10	@20
Baselines	NoDiversity	0.208	0.231	0.258	0.192	0.228	0.268	0.381	0.427	0.452
	MMR	0.221	0.232	0.261	0.192	0.228	0.269	0.394	0.436	0.467
	PLSA	0.208	0.232	0.258	0.196	0.237	0.280	0.389	0.430	0.458
SP+	IDF	0.211	0.239	0.262	0.202	0.250▲◆‡	0.284▲◇	0.387	0.428	0.463
	Imp	0.222	0.242	0.266	0.220†	0.253▲	0.286▲◇	0.404	0.443◇	0.476
	Sim	**0.237**	0.252	0.271	0.245▲◆‡	0.267▲◆‡	0.298▲◆	0.398	0.437	0.472
Cluster+	IDF	0.233	0.248	**0.274**	0.226◆	0.261▲	0.293△◇	0.404	0.436	0.469
	Imp	0.207	0.235	0.260	0.207◇†	0.251▲◆‡	0.287▲◆‡	0.401◇	0.437	0.469
	Sim	**0.237**	**0.255**△	**0.274**	**0.260**▲◆‡	**0.290**▲◆‡	**0.326**▲◆‡	**0.427**▲◆‡	**0.466**▲◆	**0.493**▲◇†

Table 4. Test performance of diversity methods trained on TREC 2009 collection based on $\alpha - nDCG\ @20$. ▲, ◆ and ‡ denote the results are significantly better than results of NoDiversity, MMR and PLSA at 0.05 level in Wilcoxon test, respectively. △, ◇ and † denote the results are significantly better than results of NoDiversity, MMR and PLSA at 0.1 level in Wilcoxon test, respectively.

Methods		Train	Test	
		TREC09	TREC10	TREC11
Baselines	NoDiversity	0.258	0.268	0.452
	MMR	0.261	0.268	0.457
	PLSA	0.258	0.278	0.443
SP+	IDF	0.262	0.281△◆	0.459
	Imp	0.266	0.273	0.448
	Sim	0.271	0.285△◇	0.445
Cluster+	IDF	**0.274**	0.273◇	0.465†
	Imp	0.260	0.279▲◆	0.467◆†
	Sim	**0.274**	**0.304**▲◆†	**0.489**△◆‡

contained in many relevant documents and ignore the information contained in few non-relevant documents. Therefore, they can extract reasonable semantic units to represent the subtopics and diversify documents. *PLSA* results may contain many non-relevant terms while *MMR* may rank a lot of non-relevant documents to the top of diversified results since it favors documents more different from selected documents.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Second, most *Cluster*, i.e., pattern cluster based, methods perform better than *SP*, i.e., single pattern based, methods. The reason is that *Cluster* can group semantically related patterns together which can better represent the subtopics. We will compare subtopics extracted using *Cluster* and *SP* in the sub-section of subtopic modeling results.

Third, the semantic similarity-based weighting, i.e., *Sim*, is the best weighting strategy that outperforms the other strategies on most collections. The better performance of *Sim* is caused by the fact that it ranks terms or subtopics based on their similarities with the query while the other weighting strategies based on only their occurrences in the collection and ignore the presence of the query.

Cluster+Sim method can significantly improve the performance over all the baseline methods on both two test collections in Table 4. Therefore, it would be the method we recommend to use in pattern-based methods.

Parameter Sensitivity

In this section, we show the impact of parameters on the diversification performances. The parameters in all the steps described in the subsection of experiment design include the segment length in the retrieval step, *min_supp* in the step of pattern extraction, K which is the number of subtopics, the number of subtopic terms in the step of subtopic modeling and λ in the diversification step. We show the result of these parameters on the TREC09 collection.

We first tune the parameters in the frequent pattern mining. We use *SP+Sim* method to tune the values of segment length and *min_supp* when fixing all other parameters. The patterns extracted with optimum values of segment length and *min_supp* will be used for all the *single pattern based methods* and *pattern cluster based methods* to ensure that all the methods use the same set of frequent patterns.

Figure 1 shows the impact of segment length on the diversification performance. The performance first increases and then decreases when we increase the segment length. The optimum segment length is 50. When the segment length is too small, the segments may split one real subtopic into several segments. Therefore, multiple extracted patterns may cover the same real subtopic and the documents covering that subtopic would be ranked higher than documents covering the subtopic contained in one pattern. When the segment length is too big, one segment may contain multiple semantic units and one extracted patterns may contain multiple subtopics. Therefore, the documents covering different real subtopics contained in the same pattern cannot be correctly diversified.

Figure 2 shows the diversification performances with different values of *min_supp*. The system achieves the best performance when *min_supp* is 4 which is a relatively small value comparing to the number of segments, i.e., 1000, used to extracted frequent patterns. The reason is that the original retrieval result contains a lot of non-relevant segments and a large value of *min_supp* would ignore the real subtopics occurring in a small number of segments. However, the frequent patterns using 4 as the value of *min_supp* include many noisy terms which would increase the difficulty of selecting subtopics from the patterns. Therefore, we can expect that the pattern-based subtopic modeling methods can perform better when the original retrieval result contains more relevant documents. Actually, the pattern based method, i.e., *Cluster+Sim*, improves the performance more significantly over the baselines on TREC10 and TREC11 collection where the original retrieval result is better as shown in Table 3.

We then use the frequent patterns extracted with the optimum values of segment length and *min_supp* in the following steps. In the rest of this section, we compare the *single pattern based methods* and the *pattern cluster based method* with *Sim* weighting function in steps of subtopic modeling and search result diversification.

Figure 3 shows the performances of the two methods with different numbers of subtopics. The optimum number of subtopics is 3 in *SP+Sim* and 2 in *Cluster+Sim*. It is reasonable

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 1. Impact of segment length on the diversification performance

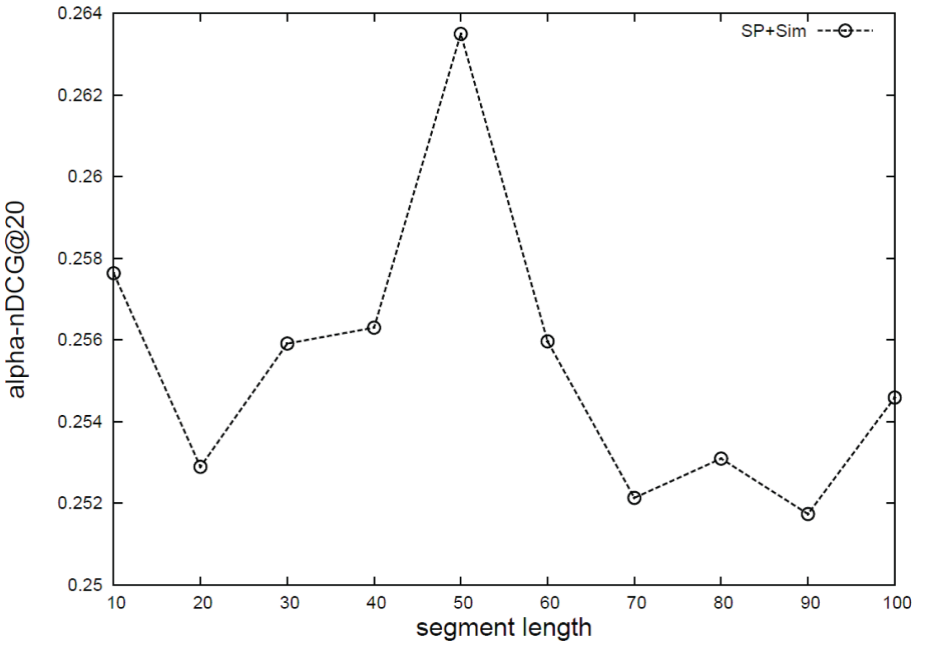
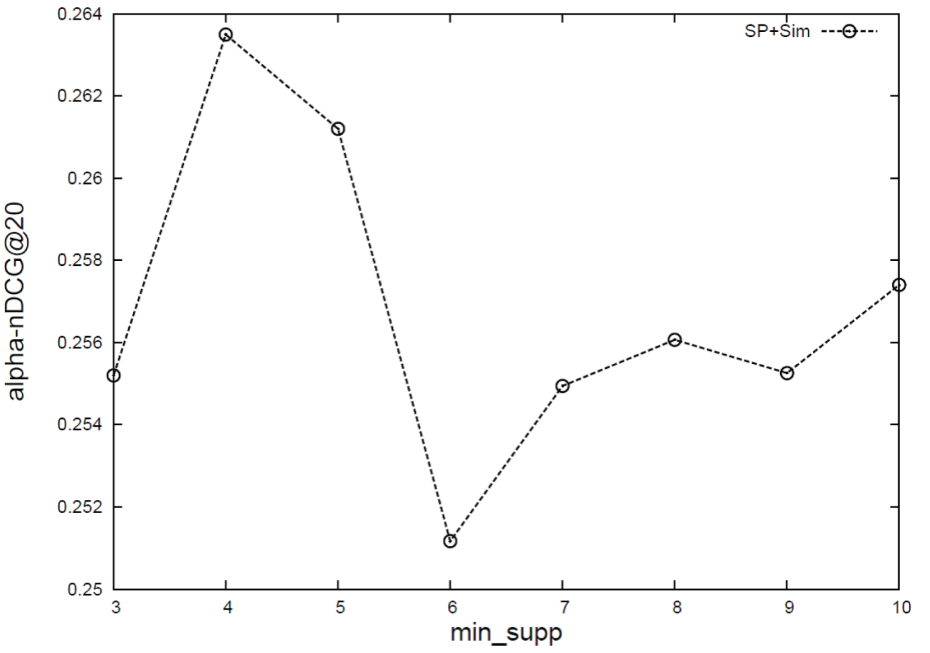
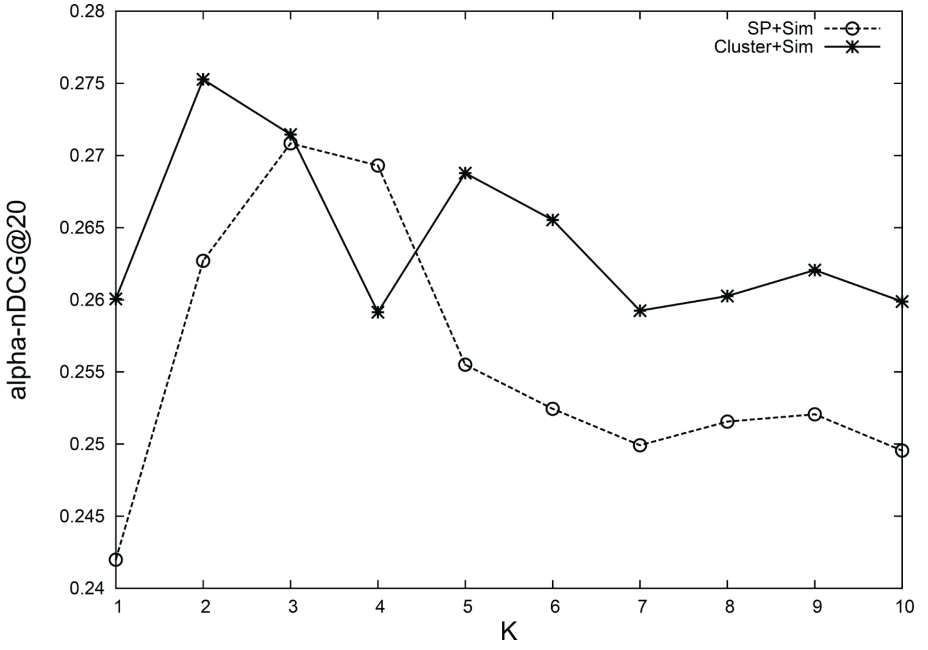


Figure 2. Impact of min_supp on the diversification performance



Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 3. Impact of K, i.e., the number of subtopics, on the diversification performances



that the optimum number of subtopics in *Cluster* is smaller than the optimum number in *SP* since pattern cluster based methods may combine multiple subtopics of *SP* into one subtopic.

Figure 4 shows the performances with different numbers of subtopic terms. In *SP+Sim*, the performance first increases and then decreases with the increase of number of subtopic terms. The trend of *Cluster+Sim* is different from *SP+Sim* and the optimum value in *Cluster+Sim* is much larger, i.e., 50, than the optimum value in *SP+Sim*. The reason is that each pattern cluster has much more terms than the single pattern. The *Sim* scores of relevant terms may not be the highest in the cluster and it is more difficult to choose relevant terms. Therefore, the *pattern cluster based method* selects more terms in each subtopic to ensure that the real subtopic terms can be included.

Figure 5 shows the impact of λ. The smaller λ is, the more the system focus on the diversity of the results. The optimum values of λ in both methods are 0.2. There are two interesting observations. (1) The performances with λ equal to 0 are worse than the performances with λ equal to 1. The reason is that the similarity between the document and subtopics cannot fully represent the relevance of the documents because the quality of the subtopics is worse than the original query. Therefore, we still need to consider the original relevance score of the document. (2)When λ is between 0 and 1, in most cases, the systems can consistently outperform the system with λ equal to 1. It means that the system performance can be improved whenever we integrate the diversification component using pattern based subtopics.

Subtopic Modeling Results

We now report the discovered subtopics using the proposed two methods. We choose query "poker tournaments" (wt09-17) as an example. Based on the judgment file, the query has six subtopics which, respectively, are "information on the world series of poker", "schedule of

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Figure 4. Impact of number of subtopic terms on the diversification performances

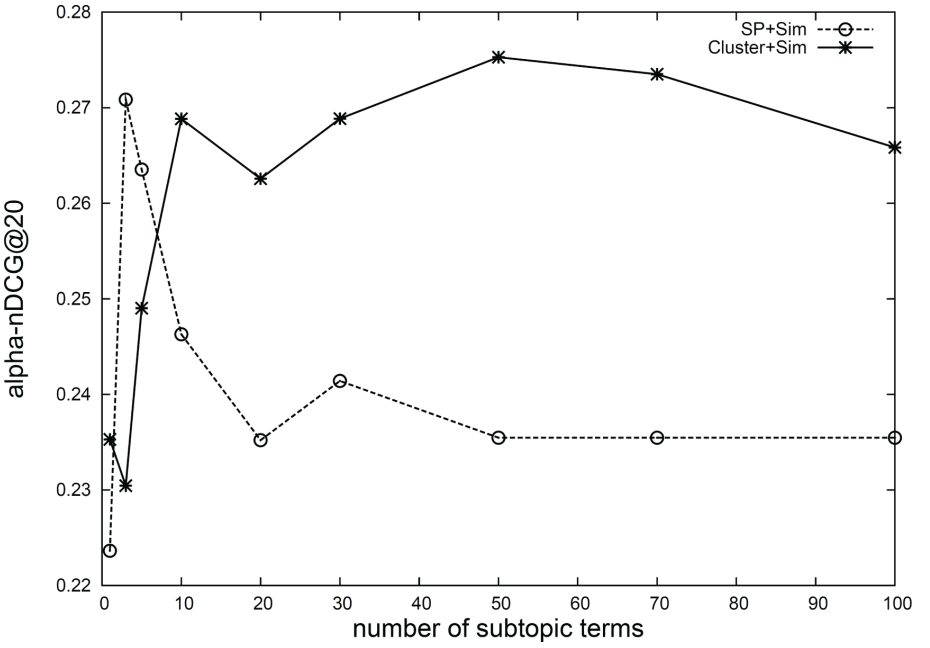
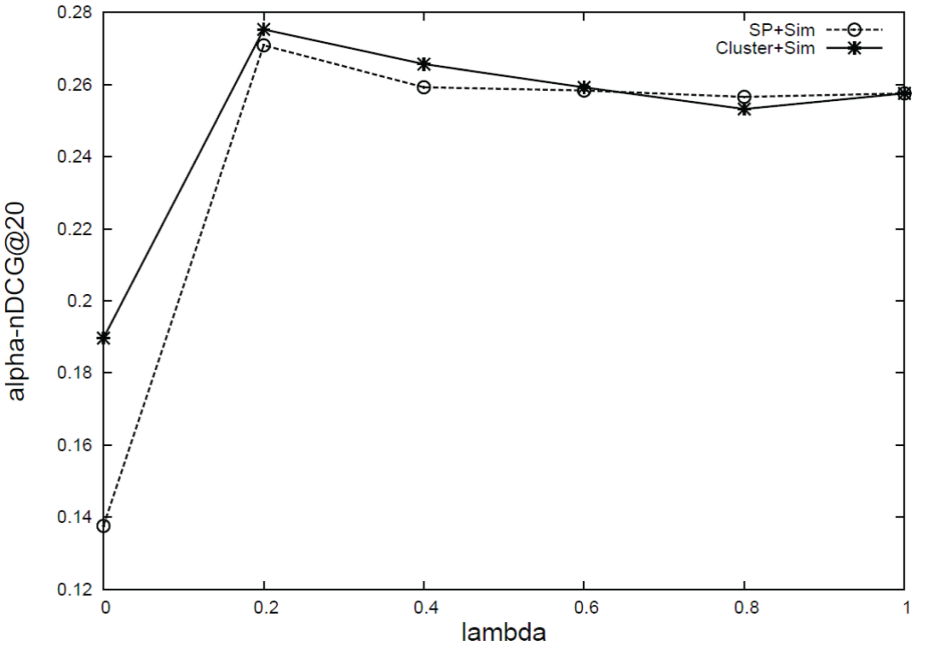


Figure 5. Impact of λ on the diversification performance



Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

poker tournaments in Las Vegas", "full tilt poker website", "schedule of poker tournaments in Atlantic City", "Texas Hold-Em tournaments" and "books on tournament poker playing".

Table 5 shows the results of the *single pattern based method*, and Table 6 shows the results of the *pattern cluster based method*. For every discovered query subtopic, we show the top 8 words in the descending order of term weights that are computed using term weighting strategies. We use strategies *Imp* and *Sim* as examples and only report their results. Based on Table 5 and Table 6, we find that all the methods can find some correct subtopic terms, such as "schedule", "texas", "las vegas", "atlantic city", etc. These results suggest that

Table 5. Subtopics discovered using single pattern based methods

SP+Imp				**SP+Sim**			
s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
Play	player	casino	player	player	player	strategy	strategy
Online	online	**texas**	online	strategy	online	freeroll	online
World	home	bonus	**world**	online	**schedule**	**wsop**	**holdem**
Room	**schedule**	**las**	**schedule**	**schedule**	**world**	satellite	hand
Freeroll	new	**vegas**	**series**	**texas**	prize	guarantee	table
Bonus	article	card	satellite	**tilt**	satellite	rakeback	odd
guarantee	hand	deposit	**tour**	rakeback	sunday	rule	software
Best	sunday	limit	sunday	**vegas**	**series**	**wpt**	best

Table 6. Subtopics discovered using pattern cluster based method

Cluster+Imp				**Cluster+Sim**			
s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
player	room	game	**world**	player	freeroll	**holdem**	**world**
strategy	freeroll	casino	**series**	strategy	room	bonus	prize
online	Site	**texas**	**tour**	online	**tilt**	**texas**	**series**
home	satellite	bonus	championship	**schedule**	**wsop**	omaha	championship
schedule	guarantee	**holdem**	prize	sunday	satellite	deposit	**tour**
new	Best	**las**	event	hand	guarantee	**vegas**	bellagio
article	Rule	**vegas**	type	rakeback	rule	hold	winner
hand	Full	city	popular	**atlantic**	pokerstar	em	stake

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

the proposed subtopic modeling methods are effective to find meaningful subtopics for a query. Moreover, we make the following two observations. First, the *pattern cluster based methods* can group terms in the same real subtopic together. For example, *Cluster+Sim* can put "holdem" in the same subtopic as the "texas" and "vegas" porker tournaments, while *SP+Sim* would assign the term to subtopics with many noisy terms. Second, it is clear that the semantic similarity-based weighting (*Sim*) is more effective to find subtopics terms compared with the term importance score weighting (*Imp*), since *SP+Sim* is able to find more meaningful subtopics than *SP+Imp*. In particular, *Sim* can find more terms relevant to the real subtopics and rank them on the top of the list because it can consider the semantic similarity between a term and the query while Imp often ignores term relations and only focuses on the term. For example, accordingly to Table 5, *SP+Sim* is able to find three very relevant terms, i.e., "wsop" (World series of poker) and "wpt" (world poker tour) and "holdem" and rank them among the top eight terms for the subtopics while *SP+Imp* is unable to do so.

CONCLUSION AND FUTURE WORK

We study the problem of search result diversification. The goal is to return a list of relevant documents covering all of the subtopics of a query while avoiding the excessive redundant information. In this paper, we propose to directly model the diversity through pattern-based subtopics. Specifically, a pattern is a semantically meaningful text unit such as a set of terms that co-occur frequently in a document collection. We first apply a maximal frequent itemset algorithm to extract the patterns from a set of retrieved documents. We then explore two ways of modeling subtopics with the extracted patterns. In the first method, we assume that a subtopic can be modeled as a single pattern and use the top ranked patterns as the query subtopics. In the second method, we assume that a subtopic is modeled as a group of related patterns and apply a profile-based clustering method to group the patterns. Given the discovered query subtopics, we re-rank the retrieval results to maximize their coverage of the subtopics.

Experiment results over the standard TREC collections show that the proposed pattern-based methods are effective in discovering subtopics and diversifying the search results.

Compared with existing studies on result diversification, the unique advantages of the proposed methods include: (1) the query subtopics are directly modeled with patterns, i.e., semantically meaningful text units; and (2) pattern-based methods allow us to focus on the important content of the documents and are more robust to the noises in the documents.

There are many interesting future research directions. First, the proposed methods use only the information from the collections. We plan to extend our methods to utilize other resources such as query logs to further improve the performance. Second, it would be interesting to apply the pattern based method to other IR problems such as query expansion and personalized search.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Number IIS-1017026. We thank the reviewers for their useful comments.

REFERENCES

Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 5-14). New York, NY: ACM.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). New York, NY: ACM.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Balog, K., Bron, M., He, J., Hofmann, K., Meij, E., & Rijke, M. … Weerkamp, W. (2009). The University of Amsterdam at TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://ilps.science.uva.nl/sites/default/files/trec2009-wn.pdf

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data* (pp. 85-93). New York, NY: ACM.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 222-229). New York, NY: ACM.

Bi, W., Yu, X., Liu, Y., Guan, F., Peng, Z., Xu, H., & Cheng, X. (2009). ICTNET at web track 2009 diversity task. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec18/papers/ictnet.WEB-DIV.pdf

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

Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing & Management*, *18*(3), 105–109. doi:10.1016/0306-4573(82)90033-4.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335-336). New York, NY: ACM.

Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1287-1296). New York, NY: ACM.

Chen, H., & Karger, D. R. (2006). Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 429 - 436). New York, NY: ACM.

Clarke, C. L. A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Overview of the TREC 2010 web track. In *Proceedings of the Nineteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec19/papers/WEB.OVERVIEW.pdf

Clarke, C. L. A., Craswell, N., Soboroff, I., & Voorhees, E. M. (2011). Overview of the TREC 2011 web track. In *Proceedings of the Twentieth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec20/papers/WEB.OVERVIEW.pdf

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley-Interscience. doi:10.1002/0471200611.

Craswell, N., Fetterly, D., Najork, M., Robertson, S., & Yilmaz, E. (2009). Microsoft Research at TREC 2009: Web and relevance feedback track. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec18/papers/microsoft.WEB.RF.pdf

Dou, Z., Chen, K., Song, R., Ma, Y., Shi, S., & Wen, J. R. (2009). Microsoft Research Asia at the web track of TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec18/papers/microsoft-asia.WEB.pdf

Fang, H., & Zhai, C. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 115-122). New York, NY: ACM.

Goffman, W. (1964). A search procedure for information retrieval. *Information Storage and Retrieval*, *2*(1), 73–78. doi:10.1016/0020-0271(64)90006-3.

Gollapudi, S., & Sharma, A. (2009). An axiomatic approach for result diversification. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 381-390). New York, NY: ACM.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of data* (pp. 1-12). New York, NY: ACM.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 289-296). San Francisco, CA: Morgan Kaufmann.

Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 120-127). New York, NY: ACM.

Li, Z., Cheng, F., Xiang, Q., Miao, J., Xue, Y., Zhu, T., et al. (2009). THUIR at TREC 2009 web track: Finding relevant and diverse results for large scale web search. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved December 9, 2012, from http://trec.nist.gov/pubs/trec18/papers/tsinghuau.WEB.pdf

Mccreadie, R., Macdonald, C., Ounis, I., Peng, J., & Santos, R. (2009). University of Glasgow at TREC 2009: Experiments with Terrier. In *Proceedings of the Eighteenth Text REtrieval Conference*. Retrieved 2012, December 9, from http://trec.nist.gov/pubs/trec18/papers/uglasgow.BLOG.ENT.MQ.RF.WEB.pdf

Mendenhall, W., Wackerly, D. D., & Schaeffer, R. L. (1990). *Mathematical statistics with applications*. Boston, MA: PWS-KENT.

Radlinski, F., Bennett, P. N., Carterette, B., & Joachims, T. (2009). Redundancy, diversity and interdependent document relevance. *ACM SIGIR Forum, 43*(2), 46-52.

Radlinski, F., & Dumais, S. (2006). Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 691-692). New York, NY: ACM.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*, 513–523. doi:10.1016/0306-4573(88)90021-0.

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 881-890). New York, NY: ACM.

Schütze, H., & Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, *33*(3), 307–318. doi:10.1016/S0306-4573(96)00068-4.

Swaminathan, A., Mathew, C. V., & Kirovski, D. (2009). Essential pages. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (pp. 173-182). Washington, DC: IEEE Computer Society.

Xue, G.-R., Dai, W., Yang, Q., & Yu, Y. (2008). Topic-bridged PLSA for cross-domain text classification. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 627-634). New York, NY: ACM.

Yan, X., Cheng, H., Han, J., & Xin, D. (2005). Summarizing itemset patterns: A profile-based approach. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 314-323). New York, NY: ACM.

Yue, Y., & Joachims, T. (2008). Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1224-1231). New York, NY: ACM.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, *12*(3), 372–390. doi:10.1109/69.846291.

Zhai, C., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 10-17). New York, NY: ACM.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 334-342). New York, NY: ACM.

Zheng, W., & Fang, H. (2011). A comparative study of search result diversification methods. *Proceedings of Diversity in Document Retrieval 2011*. Retrieved December 9, 2012, from http://www.eecis.udel.edu/~zwei/paper/ddr.pdf

Zheng, W., Xuanhui, W., Fang, H., & Cheng, H. (2012). Coverage-based search result diversification. *Journal of Information Retrieval*, *15*(5), 433–457. doi:10.1007/s10791-011-9178-4.

Zobel, J., & Moffat, A. (1998). Exploring the similarity space. *ACM SIGIR Forum, 32*(1), 18-34.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Wei Zheng is a PhD candidate in the Department of Electrical and Computer Engineering at the University of Delaware. He received his MS and BS degrees from Harbin Institute of Technology in 2006 and 2008, respectively. His PhD research is on diversifying the search results of the query to satisfy all the information needs of users in the query. He also worked on retrieval model, data mining, entity track, topic tracking, and personalization.

Hui Fang is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Delaware. Dr. Fang received the MS and PhD degree from University of Illinois at Urbana-Champaign in 2004 and 2007, respectively, and BS degree from Tsinghua University in 2001. Dr. Fang also has a broad interest in all kinds of real world applications that are related to effectively and efficiently managing large amount of text information. Dr. Fang's primary research interest is information retrieval. Dr. Fang is also interested in bioinformatics, data mining and databases.

Hong Cheng is an Assistant Professor in the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. She received her PhD degree from University of Illinois at Urbana-Champaign in 2008. Her research interests include data mining, database systems, and machine learning. She received research paper awards at ICDE'07, SIGKDD'06 and SIGKDD'05, and the certificate of recognition for the 2009 SIGKDD Doctoral Dissertation Award. She is a recipient of the 2010 Vice-Chancellor's Exemplary Teaching Award at the Chinese University of Hong Kong.

Xuanhui Wang is a research scientist in Facebook starting from 2012. Before this, Dr. Wang spent almost 3 years in Yahoo Labs and worked on online recommendation and learning to rank problems. Dr. Wang has obtained the PhD in the area of information retrieval from Department of Computer Science at University of Illinois at Urbana-Champaign in 2009 and received the Bachelor degree from University of Science and Technology of China in 2003.

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.