# An Exploration of Pattern-based Subtopic Modeling for Search Result Diversification

Wei Zheng
University of Delaware
zwei@udel.edu

Xuanhui Wang
Yahoo!
xhwang@yahoo-inc.com

Hui Fang
University of Delaware
hfang@ece.udel.edu

Hong Cheng
Chinese University of Hong Kong
hcheng@se.cuhk.edu.hk

## ABSTRACT

Traditional information retrieval models do not necessarily provide users with optimal search experience because the top ranked documents may contain the same piece of relevant information, i.e., the same subtopic of a query. The goal of search result diversification is to return search results that not only are relevant to the query but also cover different subtopics. Therefore, the subtopic modeling is an important research topic in search result diversification. In this paper, we propose a novel pattern based method to extract subtopics from retrieved documents. The basic idea is to explicitly model a query subtopic as a semantically meaningful text unit in relevant documents. We apply a frequent pattern mining algorithm to efficiently extract these text units (patterns) from retrieved documents. We then model a query subtopic with a single pattern and rank subtopics based on their similarity with the query. These pattern based subtopics are then used to diversify search results.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

## General Terms

Algorithms

## Keywords

information retrieval, diversification, subtopic, pattern

## 1. INTRODUCTION

Traditional retrieval models ignore the relations among documents, and the returned documents may contain the same piece of relevant information. Intuitively, search results covering different pieces of relevant information, i.e., query subtopics, are more desirable. The goal of search result diversification [2, 7] is to diversify the relevance documents based on their coverage of subtopics. One of the key challenges is to identify the subtopics of a query [3]. The state of the art methods include using topic modeling methods [2] to discover the subtopics from the collection and utilizing external sources such as query logs to identify the subtopics [7]. Unfortunately, the topic modeling based methods are inefficient while the external source based methods are unable to generate collection-specific query subtopics, which may lead to unsatisfying search results.

In this paper, we propose a novel pattern based subtopic modeling for result diversification, which can efficiently discover subtopics from the document collection. Specifically, we define a pattern as a semantically relevant text unit that frequently co-occur in the relevant documents. For example, for query "java", the patterns might be "programming language", "code development" and "coffee flavor". We propose to apply a maximal frequent itemset mining algorithm [1] to extract patterns. We then propose three weighting strategies to rank the patterns and select the important patterns as the subtopics. With the identified subtopics, we then use a state of the art result diversification method, i.e., $xQuAD$ [7], to diversify search results so that they are not only relevant to the query but also cover more subtopics.

## 2. PATTERN-BASED SUBTOPIC MODELING

### 2.1 Pattern Extraction

A pattern can be defined as a semantically meaningful text unit. Since the co-occurrences of terms often indicate semantic relationships between the terms [4], we define a pattern as a set of terms that satisfy the following two requirements: (1) all of the terms in a pattern need to co-occur no less than $min\_sup$ times in the retrieved document set of the query; and (2) there exists no superset in which all of the terms co-occur no less than $min\_sup$ times in retrieved documents. The second requirement avoid us being overwhelmed with small patterns with redundant information.

In fact, if we assume that a term is an item and a document is a transaction in databases, the definition of patterns is very similar to that of the maximal frequent itemsets [1]. Thus, we propose to adapt a maximal frequent itemset mining algorithm, i.e., Max-Miner proposed by Bayardo [1], for pattern extraction. Specifically, we first use query to retrieve documents based on Dirichlet prior smoothing [9]. We then construct a set-enumeration tree over all the terms in retrieved documents and perform a breadth-first search to find the patterns. We also use two pruning strategies, i.e., superset frequency pruning and subset infrequency pruning, to reduce the search space. In particular, this algorithm scales


*JCDL'11,* June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

**Table 1: Diversification performance comparison**

| | $\alpha$-nDCG | | | precision-IA | | |
|---|---|---|---|---|---|---|
| | @5 | @10 | @20 | @5 | @10 | 20 |
| NoDiversity | 0.151 | 0.183 | 0.225 | 0.069 | 0.069 | 0.069 |
| PLSA | 0.197 | 0.209 | 0.240 | 0.100 | 0.083 | 0.078 |
| Pattern+IDF | 0.187 | 0.221 | 0.256 | 0.091 | 0.088 | **0.090** |
| Pattern+Imp | 0.217 | 0.248 | **0.287** | 0.100 | 0.094 | 0.089 |
| Pattern+Sim | **0.226** | **0.252** | 0.279 | 0.104 | **0.098** | 0.087 |

**Table 2: Subtopics discovered using pattern based methods (Query: "poker tournaments")**

| Pattern+Imp | Pattern+Sim |
|---|---|
| world, room, freeroll, best | play, bonus, room, sunday |
| freeroll, site, satellite, rule | online, **schedule**, **texas**, **vegas** |
| home, **schedule**, new, sunday | **schedule**, **world**, prize, **series** |
| bonus, **las**, **vegas**, limit | **wsop**, satellite, guarantee, **wpt** |
| **world**, **schedule**, **series**, sunday | **holdem**, hand, table, best |

roughly linearly in the number of patterns and the size of document collection [1], which is clearly more efficient than the discussed term similarity based method.

## 2.2 Pattern based Subtopic Modeling

We treat each pattern as a subtopic candidate and select the most important patterns as the subtopics. We explore three weighting strategies to compute the importance of each subtopic, i.e., ***IDF*** [6], *Term importance score [8] (**Imp**)* and *Semantic similarity based weighting [4] (**Sim**)*.

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

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

$$weight_{Sim}(S) = \sum_{t \in S} \frac{\sum_{q \in Q} weight_{idf}(q) \cdot sim(q,t)}{|Q|} \quad (3)$$

where $S$ is a subtopic candidate, $df(t)$ is the number of documents containing term $t$, $N$ is the number of documents in the collection, $|Q|$ is the length of the query, and $sim(q,t)$ denotes the mutual information based term similarity [4].

# 3. EXPERIMENTS

## 3.1 Experiment Design

We evaluate the effectiveness of the proposed methods over the standard collection used for the diversity task in TREC 2009 [3]. There are 50 official topics and 50 million English-language pages in the "Category B" data set. The performance is measured using two official measures [3]: $\alpha$-nDCG and precision-IA at three depths, i.e., the top 5, top 10 and top 20 retrieved documents. We use *xQuAD* [7] to retrieve documents that not only are similar to the query but also cover more important subtopics extracted by the pattern based subtopic modeling method. We also implement two baseline methods: (1)*NoDiversity*, where we rank documents based on only relevance; and (2) *PLSA*, where we use the probabilistic latent semantic analysis algorithm [5] to extract the subtopics and *xQuAD* to diversify results. We set the number of subtopics to five for all the methods.

## 3.2 Effectiveness of Pattern-based Subtopic Modeling

Table 1 shows the performance of all the compared methods. We make the following interesting observations. (1) The pattern based subtopic modeling methods are more effective than the existing topic modeling based method, i.e., PLSA. (2) The semantic similarity-based weighting (*Sim*) is the best weighting strategy. It ranks subtopics based on their similarities with the query while the other weighting strategies ignore their relations with the query. We also compare our methods with the original *xQuAD* method [7], which identify the subtopics using query suggestions of Web search engines. The pattern based subtopic modeling method, e.g., *Pattern+Sim*, outperforms the previously studied query suggestion based method whose best performance of their method is 0.208 measured based on $\alpha$-nDCG@10.

## 3.3 Subtopic Modeling Results

We choose the query "poker tournaments" (wt09-17) as an example to report the discovered subtopics. The real subtopics of the query are: "information on the world series of poker", "schedule of 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 2 shows the top terms of extracted subtopics. Terms related to the real subtopics are highlighted. The results of *Pattern+IDF* are not reported because its performance is worse than the other two methods. The table shows that both pattern based subtopic modeling methods are effective to find terms of relevant subtopics. However, it is clear that *Sim* is more effective compared with *Imp*, because it can consider the semantic similarity between a term and the query. For example, *Sim* is able to find three very relevant terms, i.e., "wsop" (World series of poker) and "wpt" (world poker tour) and "holdem" while *Imp* is unable to do so.

# 4. CONCLUSION AND FUTURE WORK

We study the problem of subtopic modeling for search result diversification. 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; (2) the patterns can be extracted rather efficiently; and (3) pattern-based methods allow us to focus on the important content and are more robust to the noises in the documents.

# 5. REFERENCES


[1] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of SIGMOD'98*, 1998.

[2] B. Carterette and P. Chandar. Probabilistic models of novel document rankings for faceted topic retrieval. In *Proceedings of CIKM'09*, 2009.

[3] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. In *Proceedings of TREC'09*, 2009.

[4] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of SIGIR'06*, 2006.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of UAI'99*, 1999.

[6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[7] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of WWW'10*, 2010.

[8] A. Swaminathan, D. V. Mathew, and D. Kirovski. Essential pages. Technical Report MSR-TR-2008-15, 2008.

[9] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'01*, 2001.