A Study of Pattern-based Subtopic Discovery and Integration in the Web Track

Wei Zheng and Hui Fang

Department of ECE, University of Delaware

Abstract

We report our systems and experiments in the diversity task of TREC 2011 Web track. Our goal is to evaluate the effectiveness of the proposed methods for subtopic extraction and diversification steps on the large data collection. In the subtopic extraction step, we extract subtopics using both structured data, i.e., ODP, which provides high quality information and unstructured data, i.e., original retrieved documents, which contains terms effective in diversifying documents. In the diversification step, we use a coverage-based method to diversify documents based on the extracted subtopics. It has the desired properties of diversification which favors documents covering more subtopics and documents covering novel subtopics that have not been well covered by previously selected documents.

1 Introduction

The InfoLab from the ECE department at University of Delaware participated in the diversity task of TREC 2011 web track. We evalute the proposed methods on the TREC collection.

Search result diversification has attracted a lot of attention recently [3]. A commonly used strategy is to diversify documents based on the coverage of query subtopics [3, 8]. The goal of subtopic-based diversification is to maximize the coverage of query subtopics in the retrieved documents. There are two main steps in subtopic-based diversification. One step is to extract subtopics of the query and the other step is to diversify documents based on the extracted subtopics.

Existing studies of subtopic extraction used the information from structured data, i.e., taxonomy [1, 7], or unstructured data, i.e., retrieved documents [4, 10] and query suggestions [2]. The structured data provide high quality information but there is often a vocabulary gap between the taxonomy and the retrieved documents which could limit the effectiveness of the subtopics. The unstructured data contain terms that are effective in diversifying documents but also contain a lot of noisy terms. We integrate the structured data and unstructured data to extract high quality information that is effective in diversifying documents [9].

Given the extracted subtopics, we diversify documents based on their coverage of the extracted subtopics. The diversification function iteratively selects documents that are not only similar to the query but also cover subtopics that are not well covered by previously selected documents.

2 Subtopic Extraction

The first step of subtopic-based diversification is to extract subtopics of the query. We use the method that integrates the structured and unstructured data to extract subtopics effective in diversifying documents. It first separately selects documents from structured data, i.e., Open Directory Project (ODP [1]), and unstructured data, i.e., original retrieved documents. It then combines these subtopics and generates the integrated subtopics.

2.1 Subtopic Extraction in Structured Data

We use ODP as the structured data to extract subtopics. ODP is a multiple-level concept hierarchy where the nodes in the upper level are more general than those in the lower level. Therefore, it is clear that not only the content of the node but also the structure contain useful information for subtopic extraction.

The main idea of our method is that we select the most important nodes given the query as subtopics of the query [9]. The node is important if not only themselves but also their descendants are similar to the query. We compute the important score of a node s_i given the query based on the average similarity between the nodes in the sub-tree rooted at s_i and the query.

$$rel(s_i, q) = \frac{\sum_{s \in T_{s_i}} sim(s, q)}{|T_{s_i}|} \tag{1}$$

where s_i is the ith node in the hierarchy, i.e., a subtopic candidate, T_{s_i} is the sub-tree rooted at s_i and q is the query. $sim(s, q)$ is the semantic similarity between s and the query. We compute it based on the term co-occurrence information.

$$sim(s, q) = \frac{\sum_{t \in s} sim(t, q)}{|s|} \tag{2}$$

where t is a term in s and $sim(t, q)$ is the semantic similarity between the term and the query based on term co-occurrence information in the document set [5].

2.2 Subtopic Extraction in Unstructured Data

We extract subtopics from original retrieved documents of the query using the pattern based method [10]. A pattern is a semantically meaningful text unit whose terms frequently co-occur in the retrieved documents. We mine each set of terms that co-occur no less than N times as the subtopic candidates. We then compute the semantic similarity between these candidates and the query [5]. The candidates that are most similar to the query are selected as subtopics.

2.3 Subtopic Integration

We integrate the subtopics extracted from structured data and the subtopics extracted from unstructured data to generate integrated subtopics.

The task of subtopic integration in this section is that, given the K subtopics extracted from the structured data and K subtopics extracted from documents, we combine them into K integrated subtopics where each subtopic contains M terms. Since the final goal of search result diversification is to diversify documents, we propose to use the subtopics extracted from structured data, containing high-quality information, to guide selection of integrated subtopic terms from subtopics extracted from documents. Specifically, in each query, we first propose to connect each subtopic of structured data with a subtopic of documents based on their semantic similarity:

$$s_i = \arg\max_{s \in S} sim(s, s'_i) \tag{3}$$

[1] http://www.dmoz.org/

where S is the set of subtopics extracted from the structured data, s'_i is the ith subtopic extracted from the documents, $sim(s, s'_i)$ is the semantic similarity between the subtopic of structured data and the subtopic of documents, and s_i is the subtopic of structured data assigned to s'_i. We assume the connection between these subtopics is 1 to 1, in order to simplify the problem. We leave other methods of connection for our future work.

For each subtopic extracted from the documents, we then select terms based on their semantic similarity to the connected subtopic of structured data [5]. The selected terms from each subtopic would form a new integrated subtopic that utilizes the information from both structured data and documents. These integrated subtopics are generated by the guidance of subtopics extracted from structured data, so they often contain the information of higher quality. Moreover, their terms are extracted from the clusters of documents, so they could solve the problem of vocabulary mismatch and are more effective in diversifying documents.

3 Diversification Function

Given the extracted subtopics, we then use the diversification function to diversifying documents based on their coverage of subtopics.

The diversification problem is often formulated as an optimization problem that aims to maximize an objective function related to both the *relevance* and *diversity* of the search results. Formally, given a query q, a set of documents $\mathcal{D}$ and an integer k, the goal is to find D, i.e., a subset with k documents from the document set, that can maximize the following objective function:

$$G(D) = \lambda \sum_{d \in D} rel(q, d) + (1 - \lambda) \sum_{s \in S} weight(s, q) \cdot cov(s, D), \tag{4}$$

where $D \subseteq \mathcal{D}$, $rel(q, d)$ is the relevance score of the document in the query, S is the subtopic set of the query, $weight(s, q)$ is the weight of the subtopic in the query and $cov(s, D)$ is the coverage of the document set D on the subtopic s.

We can define $cov(s, D)$ as follows:

$$cov(s, D) = 1 - (1 - \sum_{d \in D} cov(s, d))^2 \tag{5}$$

where $cov(s, d)$ is the coverage of the document on the subtopic. We use the probability scores $P(d|q)$, $P(s|q)$ and $P(d|s)$ as $rel(q, d)$, $weight(s, q)$ and $cov(s, d)$, respectively.

We use a greedy algorithm that iteratively selects documents that has the largest gain of the optimization function.

$$d^* \quad = \quad \arg \max_{d \in \mathcal{D} \setminus D} (G(D \cup \{d\}) - G(D)) \tag{6}$$

Therefore, we can get the final diversification function:

$$d^* = \arg \max_{d}((1 - \lambda)P(d|q) + \lambda \sum_{s \in S}(P(s|q)P(d|s)(2 - 2 \cdot \sum_{d' \in D} P(d'|s) - P(d|s))), \tag{7}$$

where $P(d|q)$ and $P(d|s)$ measures the relevance scores of d with respect to q and s, $P(s|q)$ denotes the importance of s given q, and λ is a parameter that controls the balance of the relevance and diversity scores. We can see that the documents are iteratively selected based on not only their relevance to the query but also their ability of covering more subtopics that are not covered by previously selected documents. It has the following properties: (1) Diminishing return. If the document d covers the subtopics that have been better covered by previously selected documents in D, the gain of selecting this document should be smaller. (2) Favoring diversity. It favors documents that cover more subtopics. (3) Novelty emphasis. It favors documents covering the subtopic that is not well covered by the previously selected documents.

Table 1: The performances of submitted runs based on *ERR-IA*

Methods	ERR-IA@5	ERR-IA@10	ERR-IA@20
UDPattern	0.3276	0.3449	0.3573
UDCombine1	0.3256	0.3451	0.3526
UDCombine2	**0.3493**	**0.3664**	**0.3747**

4 Experiment Results

We submitted three runs in the diversity task of web track. All of them are based on the Category B collection of ClueWeb09 corpus. They use the diversification methods described in Section 3. They mainly differ in the subtopic extraction step.

1. UDPattern. It extracts subtopics directly from unstructured data using the patter based method.
2. UDCombine1. It separately extracts subtopics from ODP as described in Section 2.1 and from documents using PLSA [6]. It then integrates these subtopics as described in Section 2.3.
3. UDCombine2. It separately extracts subtopics from ODP and from documents using the pattern based method. It then integrates these subtopics as described in Section 2.3.

Table 1 lists the results of our submitted runs. We can see that *UDCombine2* performs best. It shows the effectiveness of integrating subtopics extracted from and subtopics extracted from documents using pattern based method. We can also see that the performance of *UDPattern* is similar to the performance of *UDCombine1*. It shows that the pattern based method is effective in extracting subtopics from unstructured data.

5 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number IIS-1017026. We thank TREC organizers for the provided data collection.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of WSDM'09*, 2009.

[2] K. Balog, M. Bron, J. He, K. Hofmann, E. Meij, M. de Rijke, M. Tsagkias, and W. Weerkamp. The university of amesterdam at trec 2009. In *Proceedings of TREC'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] Z. Dou, K. Chen, R. Song, Y. Ma, S. Shi, and J.-R. Wen. Microsoft research asia at the web track of trec 2009. In *Proceedings of TREC'09*, 2009.

[5] H. Fang and C. Zhai. Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In *SIGIR*, 2006.

[6] P. Lubell-Doughtie and K. Hofmann. Improving result diversity using probabilistic latent semantic analysis. In *Proceedings of DIR'11*, 2011.

[7] R. L. T. Santos, C. Macdonald, and I. Ounis. Selectively Diversifying Web Search Results. In *CIKM*, 2010.

[8] R. L. T. Santos, J. Peng, C. Macdonald, and I. Ounis. Explicit search result diversification through sub-queries. In *Proceedings of ECIR'10*, 2010.

[9] W. Zheng, H. Fang, C. Yao, and M. Wang. Search result diversication for enterprise data. In *Proceedings of CIKM'11*, 2011.

[10] W. Zheng, X. Wang, H. Fang, and H. Cheng. An exploration of pattern-based subtopic modeling for search result diversification. In *Proceedings of JCDL'11*, 2011.