Exploiting Ontologies for Search Result Diversification

Wei Zheng and Hui Fang

Department of ECE, University of Delaware

Abstract

We report our systems and experimental results in the diversity task of web track 2012. Our goal is to exploit the structured data, i.e., the ontologies, as well as unstructured data for search result diversification. We use two strategies in the diversification systems. The first strategy combines the ontology and unstructured data to extract integrated subtopics. It then uses the coverage based diversification function to diversify documents based on the integrated subtopics. The second strategy exploits the structure information in the ontology for diversification. We use a structural diversification to diversify documents based on the structural relationships of their subtopics in the ontology.

1 Introduction

The InfoLab from the ECE department at University of Delaware participated in the diversity task of TREC 2012 web track. We evaluate our methods on Category B collection of ClueWeb09 corpus.

Search result diversification mainly has two steps, i.e., subtopic extraction and document diversification [2, 5, 8]. Most studies either use the unstructured data [4, 8] or just the content of structured data [1]. However, they ignore the relationships between the structured and unstructured data, and the structure information in structured data, i.e., the ontology. Therefore, we use two strategies to exploit the content and structure of the ontology.

The first strategy combines the content of the ontology and the unstructured data, i.e., documents [8]. The ontology contains high quality information while unstructured data contains information effective in diversifying documents. It would be more effective to extract subtopics for diversification when using both of them than using one of them. Therefore, we first separately extract subtopic candidates from structured and unstructured data. We then combine these candidates to generate the integrated subtopics. The integrated subtopics are then used to diversify documents using the coverage based diversification method.

The second strategy exploits the structure of the ontology for diversification [6]. We first assign each document to the most similar node in the hierarchy of the ontology. Two nodes have a lot of overlapped information if they are very close to each other in the hierarchy. We therefore use the structural diversification function to compute the diversity of each document based on the position of its node. A document has high diversity score if its assigned node is structurally far away from the nodes of selected documents. It then iteratively selects the document that is relevant to the query and structurally diversity to the selected documents.

2 Integration of Ontologies and Unstructured Data

In this section, we use the subtopic extraction method proposed by Zheng et. al. [7] to extract subtopics integrating structured data with unstructured data. The extracted subtopics can then be used to diversify documents based on any state of the art method.

We independently use the ODP [1] and DBPedia [2] as the ontologies and combine each of them with unstructured data. Each of them has a multiple-level hierarchy. The nodes on higher levels contain more general information while nodes on lower levels contain more specific information.

The system first extracts K subtopic candidates from the ontology. We directly use the query to find the most similar nodes on the hierarchy as subtopics of the query [7]:

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

where $rel(s,q)$ is the relevance score of the node s given the query q, $|T_s|$ is the number of nodes in the subtree rooted at s including s itself, s_i is the node in the subtree and $sim(s_i, q)$ is the semantically similarity between s_i and q [3]. It assumes that a node is more relevant to the query if both the general information in itself and the detailed information in its subtree are similar to the query. K non-overlapped nodes with the largest relevance score are selected as subtopic candidates.

The system uses PLSA to extract K subtopic candidates from the unstructured data [7]. It applies PLSA algorithm to generate K clusters in the top-ranked documents. We then assign each term to the cluster that most likely generates the term. Each cluster is a subtopic candidate of the unstructured data.

The system then integrates the subtopic candidates of ontology with those of unstructured data. It connects each subtopic candidate from ontology with the most semantically similar subtopic candidate from unstructured data. We assume that the connection between the subtopic candidates is 1 to 1. Each pair of connected subtopic candidates is an integrated subtopic. We then select the subtopic terms from the PLSA subtopic, which are most semantically similar to the connected subtopic candidates of ontology.

3 Ontology based Structural Diversification

This strategy directly uses documents to find the most similar nodes in the ontology as subtopics and uses the relative positions of the subtopics to measure the diversity of the documents [6].

In search result diversification, it is critical to measure the overlapping among the subtopics covered by different documents. The ontology is a good resource for that task since its structural reveals the overlapping relationships of different nodes. We therefore use the structural diversification function that iteratively selects a document at each position with the highest score [6]:

$$Score(q,d,D') = \lambda \cdot P(d|q) + (1-\lambda) \cdot \sum_{s \in S(q)} [P(s|q) \cdot P(d|s) \cdot \prod_{d' \in D'} (1 - SubCov(d', s))], \quad (2)$$

where $Score(q,d,D')$ is the score of the document d given the query q and the set of selected documents D', λ is the parameter balancing relevance and diversity, $P(d|q)$ is the relevance of the document given the query, $S(q)$ is the set of query subtopics, $P(s|q)$ is the relevance

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

[2] http://www.dbpedia.org

of the subtopic given the query, $P(d|s)$ is the relevance of the document given the subtopic, and $SubCov(d', s)$ measures how much information from subtopic s has been covered by selected document d'. We assume that each subtopic is dependent on each other and incorporate subtopic similarity into the function:

$$SubCov(d', s) = P(d'|s') \cdot \frac{\varphi(s'|s)}{\sum_{s_i \in S(q)} \varphi(s_i|s)}, \tag{3}$$

where $S(d')$ is a set of query subtopics that are relevant to d', $P(d'|s')$ measures the relevance of d' given s', the other part measures the overlapping between the selected subtopic s' and current subtopic s, $\varphi(s'|s)$ is the structural similarity of s' to s based on their positions on the hierarchy.

There are two types of edges going from s to s'. One type includes the edges going from the children UP to the parents. The other type includes edges going from the parents $DOWN$ to the children. The method uses both types of edges to compute the similarity [6]:

$$\varphi(s'|s) = \frac{\frac{2}{3}}{1 + |UP(s \rightarrow s')|} + \frac{\frac{1}{3}}{1 + |DOWN(s \rightarrow s')|}, \tag{4}$$

where $UP(s \rightarrow s')$ and $DOWN(s \rightarrow s')$ are the sets of UP edges in the edges going from s to s' and the sets of $DOWN$ edges, respectively.

The structural similarity in Equation (4) was derived based on three constraints describing the relationships between the desired similarity of subtopics and their positions in the hierarchy [6]. The idea is that two subtopics are more similar if they are structural closer to each other in the hierarchy.

1. The similarity of a subtopic to itself is not smaller than the similarity of it to any other subtopics.

2. The similarity of a subtopic's ancestor to the subtopic is not smaller than the similarity of the ancestor's ancestor to the subtopic.

3. The similarity of a subtopic's any descendant to the subtopic is not smaller than the similarity of its any ancestor to the subtopic.

We therefore can directly diversify documents when integrating Equations (3) and (4) into Equation (2).

4 Experiments

We use the Category B collection of ClueWeb09 corpus in the diversity task of web track. We submitted three runs using the strategies described in Sections 2 and 3:

1. UDInfoDivC1. It first extracts the subtopics using integrated unstructured data, i.e., documents, and the ontology, i.e., ODP, as described in Section 2. It then diversifies documents using a coverage-based diversification function, i.e., SQR [8]. The diversification function iteratively selects the documents that are more similar to the query and cover more novel subtopics that have not been well-covered by selected documents.

2. UDInfoDivC2. It uses DBPedia as the ontology and integrates it with documents to extract subtopics. It then diversifies documents using SQR.

3. UDInfoDivSt. It uses the structural diversification method as describe in Section 3 to diversify documents with DBPedia as the ontology.

Table 1: The diversification performances of submitted runs on Category B collection

Methods	ERR-IA@20	α-nDCG@20	NRBP
UDInfoDivC1	0.269	0.384	0.211
UDInfoDivC2	0.264	0.379	0.207
UDInfoDivSt	**0.300**	**0.420**	**0.241**

Table 1 shows the performances of our submitted runs. We can see that the structural diversification method performs best in the three runs. UDInfoDivSt uses both the content information and the structural information in the ontology while the other two runs focus only on the content information. It shows the importance of the structural information in the ontology to search result diversification.

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] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 Web Track. In *Proceedings of TREC'11*, 2010.

[3] H. Fang and C. Zhai. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proceedings of SIGIR'05*, 2005.

[4] P. Lubell-Doughtie and K. Hofmann. Improving Result Diversity using Probabilistic Latent Semantic Analysis. In *Proceedings of DIR'11*, 2011.

[5] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of WWW*, 2010.

[6] W. Zheng, H. Fang, and C. Yao. Exploiting concept hierarchy for result diversification. In *Proceedings of CIKM'12*, 2012.

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

[8] W. Zheng, X. Wang, H. Fang, and H. Cheng. Coverage-based search result diversification. *Journal of Information Retrieval*, 15(5):433–457, 2012.