Automatic Annotation of Environmental Metadata Using Topic Similarity
Suppawong Tuarab1, Line C. Pouchard2, Natasha Noy3, Jeffery S. Horsburgh4, and Giri Palanisamy2
Pennsylvania State University, University Park, PA; 2Oak Ridge National Laboratory, Oak Ridge, TN; 3Stanford University, Stanford, CA; 4Utah State University, Logan, UT

The rapid growth of diverse data types and greater volumes available to environmental sciences prompts the scientists to seek knowledge in data from multiple places, times, and scales. To facilitate such need, ONEMercury has recently been implemented as part of the DataONE project to serve as a portal for accessing environmental and observational data across the globe. ONEMercury harvests metadata from the data hosted by multiple repositories and makes it searchable. However, harvested metadata records sometimes are poorly annotated or lacking meaningful keywords, and hence would unlikely be retrieved during the search process. In this paper, we develop an algorithm for automatic metadata annotation. We transform the problem into a tag recommendation problem, and propose a score propagation style algorithm for tag recommendation. Our experiments on four data sets of environmental metadata records not only show great promises on the performance of our method, but also shed light on the different nature of the data sets.

Problem:
Linking data from heterogeneous sources always has a cost. One of the biggest problems that ONEMercury is facing is the lack of annotation in the harvested metadata records. Poorly annotated metadata records tend to be missed during the search process, as they lack meaningful keywords. Furthermore, such records would not be compatible with the advance mode offered by ONEMercury as it requires the metadata records be semantically annotated with keywords from the keyword library. The explosion of the amount of metadata records harvested from an increasing number of data repositories makes it even impossible to annotate the harvested manually by hand, urging the need for a tool capable of automatically annotating poorly curated metadata records.

Proposed Solution:
We transform the problem into the tag recommendation problem. We propose a novel score propagation technique for tag recommendation. Given a document query q, we first calculate the similarity score between the query and each document in the source. The score then is propagated to the tags of each document in the source. Tags then are ranked by the scores, and the top K tags are returned for recommendation. We propose two different measures for computing the similarity between two documents: term-frequency inverse document-frequency based (TFIDF) and Topic Model Based (TM).

Document Similarity Measures:
We explore two different document similarity measures when computing the similarity between the document query and the documents in the source.

TFIDF Based:
The first measure relies on the term frequency-inverse document frequency. In order to compute the document frequency, all the documents in the source need to first be indexed. We use LingPipe to perform the indexing and calculating the TFIDF based similarity.

Topic Model (TM) Based:
The second document similarity measure utilizes topic distributions of the documents. Hence the training process involves modeling topics from the source using LDA (Latent Dirichlet Allocation) algorithm. We use Stanford Topic Modeling Toolbox with the collapsed variational Bayes approximation to identify topics in the source documents. For each document we generate uni-grams, bi-grams, and tri-grams, and combine them to represent the textual content of the document. The algorithm takes two input parameters: the number of topics to be identified and the maximum number of the training iterations. After some experiments on varying the two parameters we fix them at 300 and 1,000 respectively. We use Cosine similarity to compute the topic similarity between two documents.

Evaluation:
We evaluate our methods using the tag prediction protocol. We artificially create a test query document by removing the tags from an annotated document. Our task is to predict the removed tags. We use document wise 10 fold cross validation to evaluate our schemes with 5 evaluation metrics: Precision, Recall, F1, MRR (Mean Reciprocal Rank), and BPR (Binary Preference). These evaluation metrics are extensively used together to evaluate recommendation systems.

Discussion:
The TM based approach obviously outperforms the TFIDF based approach in DAAC, DRID4, and KNB data sets. The performance of the TM based approach is dominant when applied to the KNB data set, as seen in the precision vs recall graph, where the curve forms the shape close to the ideal precision vs recall curve. The comparison, however, is not dominant on the TreeBASE data set. Actually, both algorithms perform very poorly on the TreeBASE data set. We hypothesize that this is because the TreeBASE documents are very sparse and have very few tags. From our statistics, each document in the TreeBASE data set has only 11 words and only 0.7 tags on average. Such sparse texts lead to weak relationship when finding textual similar documents in the TFIDF approach, and the poor quality of the topic model used by the TM based approach. The small number of tags per document makes it even harder to predict the right tags. We note that, though overall the TM based approach recommends better quality of tags, the training times take significantly longer than those of the TFIDF based approach.

References: