Abstract

With tremendous increase in the amount of digital data available educators are forced to author content for learning and teaching for use in their classes. With that there has emerged a need to facilitate automatic discovery of learning resources from the World Wide Web. In this work, we present a novel approach for discovering content from the web for e-learning. We argue that for an e-learning scenario, retrieval of the redundant content from the web is a serious problem to be addressed as it does not satisfy the requirements of a typical learner. Furthermore, the content retrieved should cover all topics as in his syllabus. Sense-disambiguation should be performed during information retrieval from the web so that it corresponds to the learner’s actual domain of interest. This work presents a domain ontology based re-querying approach for query expansion to discover content from open corpus sources. We use the Latent Dirichlet Allocation Model for unsupervised classification of document
segments to aid students and educators. Having identified the topics at the granularity of document segments in an unsupervised fashion, we state that internal topic transitions in a resource retrieved from the web can be exploited for providing relevant and personalized content. In addition to this, we propose a re-ranking scheme for ordering results from search engines to maximize topic coverage and minimize redundancy among retrieved results. We also evaluate the effectiveness of our proposed method for information retrieval and show that our work results in greater coverage of topics from the web without redundancy.

Reference

- Apache Lucene - a free/open source information retrieval software library, originally created in Java. It is released under the Apache Software License.
- Nutch an open source search engine based on Lucene at nutch.apache.org.
- Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry, 1999 , “The PageRank citation ranking: Bringing order to the Web”.
- Search engine available at www.google.com
Santander, Cantabria, Spain.
- Protégé ontology editor available at http://protege.stanford.edu/

**Index Terms**

Computer Science  
Artificial Intelligence

**Key words**

Latent Dirichlet Allocation
Re-ranking

Topic coverage
Reudundancy