ABSTRACT

Text mining deals with the automated annotation of texts and the extraction of facts from textual data for subsequent analysis. Such texts range from short articles and abstracts to large documents, for instance, scientific articles and web pages. This thesis focuses on different key problems in Geological text mining: named entity recognition, named entity disambiguation and relationship extraction. The goal of these text mining tasks is to develop intelligent and efficient search engines for geology researchers. The objective of such systems is to take advantage of any locational cues present in the documents and queries.

First, a corpus containing documents pertaining to geology of the Indian subcontinent is created in this work. Comparative study of the performance of different information retrieval models is performed on a geological text corpus. Most of the popular retrieval models are extensively compared through experiments.

Subsequently, named entity recognition in the geological corpus has been performed. Named entity recognition identifies mentions of geological entities, for example, country, state, city, region, mountain, island, waterbodies, village, mineral, year organization, measures, person, time, fault and rock in a text. Here recognition of location names is focussed. Conditional Random Fields with diverse lexical and semantic features has been used to recognize and classify different geological entities in text. A new named entity tagset has been developed for annotation of the corpus used or training the classifier.

Next, novel techniques to automatically resolve place name ambiguities such as reference ambiguity and referent ambiguity are proposed. The disambiguation algorithm is based on the methods for generating co-occurrence models (co-occurrence matrix and co-occurrence graph) from the Geological corpus. The co-occurrence models are used for mining co-occurrences of place names from the Geological corpus i.e., it represents mutual relationships between place names. In order to resolve place name ambiguities, a co-occurrence graph based disambiguation algorithm is proposed.

Finally, identification of relations as interactions and associations between location-location entities is done. Conditional Random Fields as well as sequence kernels are used for extracting relations between entities from a geological text corpus. The extension of framework of Conditional Random Fields towards the annotation of different semantic relations from text has been applied to the geological domain. The model is quite general and can be expanded to handle geological entities and relation types. In sequence kernels, common subsequences in strings of sentences are identified as relations and are automatically annotated. This thesis also provides test results which verify that different text mining tasks improves retrieval performance.

Each of the above techniques are used to enhance the performance of retrieval engines on the geological corpus. Often a query expansion based technique is used for this. We have found through extensive experiments that each of the above techniques are successful in improving retrieval performance. The over all contribution of the thesis is in study of several text mining techniques important for geological documents. The techniques help to develop intelligent search techniques for scientific researchers. This is the first work which has been done in geological information retrieval.