

AUTOMATIC METHODS TO DISAMBIGUATE GEOSPATIAL QUERIES

Carolyn Hafernik

Today an unprecedented amount of digital information is available, but locating information of interest can be difficult. Information Retrieval (IR) is the area of Computer Science that aims to locate information by automatically organizing, storing, and managing it. IR systems search large databases, separating non-relevant items from the relevant items, and return a list of the documents likely to match the information need as described by the user's query. Geographical Information Retrieval (GIR) aims to develop IR systems with spatial awareness and exploit geospatial information to improve retrieval effectiveness. This research explores GIR, as in GeoCLEF [1], and uses geospatial information to automatically disambiguate geospatial terms. The hypothesis is that automatically disambiguating geospatial terms will improve retrieval effectiveness.

Two challenges to IR and GIR are language ambiguity and improving retrieval automatically, instead of manually, by query modification. Language ambiguity can be problematic for both queries and documents because there are many ways to describe concepts or ideas. Separate documents describe the same information differently. Similarly, independent users looking for the same information may use different words in their queries. A query describing a concept in one way will fail to retrieve relevant documents that use different vocabulary. Users often fail to precisely specify information they require, which may cause the system to miss some important relevant documents. Manual approaches to query modification require a person to determine other concepts and words which not only better describe the needed information, but that others may have chosen for the same topic. This is not realistic in a real world situation. Often there is not enough time for individuals to modify the queries themselves and they might not know how to improve queries. Thus automatic methods, which can be done without a human are needed in order for IR or GIR methods to be practical.

This work aims to exploit geospatial information in queries to improve retrieval by automatically disambiguating geospatial terms within the queries using outside geospatial knowledge gathered from the internet, including city names, countries, regions, parts of countries and location information. Our approach combines simple linguistic analysis with query modification via the addition of geospatial information. Geospatial terms were chosen in several different ways. First, terms were added from retrieved documents assumed to be relevant [2]. Another method gave higher weight to more important query words. A third procedure added terms selected from a geographic thesaurus. Finally, attempts were made to perform spatial disambiguation by using longitude and latitude to infer an upper bound on distance terms like "near."

[1] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, Vivian Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. http://www.clef-campaign.org/2005/working_notes/CLEF2005WN-Contents.html, 2005.

[2] J. J. Rocchio, jr. Relevance Feedback in Information Retrieval. In Gerard Salton (ed.) *The SMART Retrieval System, experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc, 1971.