Geographic Information Retrieval (GIR) Ranking Methods for Digital Libraries

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@sherlock.berkeley.edu

Patricia Frontiera
College of Environmental Design
University of California, Berkeley
Berkeley, California, USA, 94720-1839
pattyf@regis.berkeley.edu

ABSTRACT

This demo will presents results from an evaluation of algorithms for ranking results by probability of relevance for Geographic Information Retrieval (GIR) applications. We will demonstrate an algorithm for GIR ranking based on logistic regression from samples of the test collection. We also show the effects of different representations of the geographic regions being searched, including minimum bounding rectangles, convex hulls, and complex polygons.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*retrieval models, search process*; H.2.8 [**Database Management**]: Database Applications—*Spatial Databases and GIS*

General Terms: Algorithms, Performance, Design

Keywords: Geographic Information Retrieval

1. EXTENDED ABSTRACT

Digital Libraries today range from small specialized collections of information on specific topics to very large-scale repositories covering a wide range of topics. With increasing frequency this data will include geospatial metadata describing the geographic or spatial extents of the data, or the data itself will contain geographically referenced information ranging from place names in texts, or in the metadata describing the objects, to geographic coordinates explicitly recorded in the data.

Simply stated, most of the objects in digital libraries are, to a greater or lesser extent, about or related to particular places on or near the surface of the Earth. However, current metadata representations of geographic characteristics and the uses of these metadata are problematic. Most retrieval systems for Digital Libraries, even those tailored to geographic information, provide only nascent approaches to the technologically and conceptually difficult challenges of GIR. Much of the problem is rooted in the geospatial metadata used by these systems to index and access geographic data. The primary issues are:

Lexical: Geographic metadata typically lack the rich and well understood textual clues, such as keywords and titles, and descriptive information, that are the primary inputs to current information retrieval and management methods.

Spatial: Geographic metadata do not sufficiently capture the geospatial characteristics of geographic data. Moreover, information retrieval systems, even those designed for geographic data, don't leverage the geospatial characteristics currently encoded in metadata to support information retrieval and management tasks.

In this demonstration we will show how geospatial metadata can be exploited to provide effective and accurate ranked retrieval of geospatial information, using retrieval algorithms based on logistic regression with weighting coefficients estimated from a set of training data.

In this demonstration we will show:

1. How graphical geospatial query specifications can be used to obtain the "best-match" sets of geospatial data.
2. How different representations of the underlying data extents, including minimum bounding rectangles and convex hulls, compare to complex polygon representations in retrieval.
3. How other characteristics, such as contextual geographic information, can be combined with knowledge of the query and candidate regions to improve retrieval effectiveness.
4. How online gazetteers can be used to apply geographic retrieval to texts.

The demonstration will show real-time live searches and geographic displays to illustrate the algorithms and methods described.

2. ACKNOWLEDGMENTS

Work on geographic information retrieval for the Cheshire II system was supported in part by the National Science Foundation and Joint Information Systems Committee(U.K) under the *International Digital Libraries Program* award #IIS-9975164. Addition support was provided by a grant from the Institute of Museum and Library Services (IMLS) entitled "Going Places in the Catalog".

JCDL '04 June 7–11, 2004, Tucson, Arizona, U.S.A.
ACM 1-58113-513-6/04/0006.