Geographic Information Retrieval (GIR): Searching Where and What

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

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

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

Information retrieval systems in operation today for applications ranging from Digital Libraries to Web Search make very little use of two major dimensions of the data being searched: location and time. For many applications these dimensions can provide an intuitive and understandable visualization of search constraints and results. However, current metadata representations of geographic characteristics and the uses of these metadata are problematic. Most retrieval and database, even those tailored to geographic information, provide only nascent approaches to the technologically and conceptually difficult challenges of Geographic Information Retrieval (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 database 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 explicit geospatial metadata and how *inferred* geographic information from texts can be exploited to provide effective and accurate ranked retrieval of geospatial information and relevant text documents using retrieval algorithms based on logistic regression with weighting coefficients estimated from a set of training data.

We will present a system that combines conventional probabilistic algorithms for text retrieval with algorithms for estimating probability of relevance for geographic spaces. We will also demonstrate our algorithm for GIR ranking that estimates probability of relevance based on a weighted set of parameters where the weights were derived using logistic regression from samples of a test collection.

In this demonstration we will show:

1. How graphical geospatial query specifications can be used to obtain sets of geospatial data ranked by probability of relevance.

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".

SIGIR '04 July 25–29, 2004, Sheffield, South Yorkshire, UK.
Copyright 2004 ACM 1-58113-881-4/04/0007 ...$5.00.