Geographic Information Retrieval and Digital Libraries

Ray R. Larson

School of Information
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@sims.berkeley.edu

Abstract. In this demonstration we will examine the effectiveness of Geographic Information Retrieval (GIR) methods in digital library interfaces. We will show how various types of information may benefit from explicit geographic search, and where text-based place name search may be sufficient. We will also show how implicit geographic search (or geographic browsing) can be used to dynamically generate geographic searches in geographic interfaces like Google Earth. In this demostration we will show the algorithms used for Geographic search and how these may be combined with text search. In addition we will show results from the GeoCLEF IR evaluation for text-based search.

1 Geographic Information Retrieval

The goal of Geographic Information Retrieval (GIR) is to retrieve *relevant* information resources in response to queries with geographic constraints. GIR implies that the indexing and retrieval of objects in a digital library collection takes into account some form of georeferencing[2], and may use various forms of geographical proximity, containment, or other spatial relations in estimating or predicting relevance. Systems that provide searches using GIR methods, including geographic digital libraries, and location-aware web search engines, are based on a collection of georeferenced information resources and methods to spatially search these resources with geographic location as part of their search specifications.

Information resources in digital library collections can be considered georeferenced if they are spatially indexed by one or more regions or points on the surface of the Earth, where the specific locations of these regions are encoded using spatial coordinates directly (*geometrically*), or indirectly by toponyms (*place names*).

One common approach in digital libraries has been to use place names as a geographical search surrogate. However, place names have well-documented lexical and geographical problems [3]. Lexical problems include lack of uniqueness, variant names or spellings, and name changes. Geographical problems include boundaries that change over time and geographic features or areas without known place names. While geographic coordinates provide can an unambiguous and persistent method for locating geographic areas or features, they also present

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 461–464, 2009.

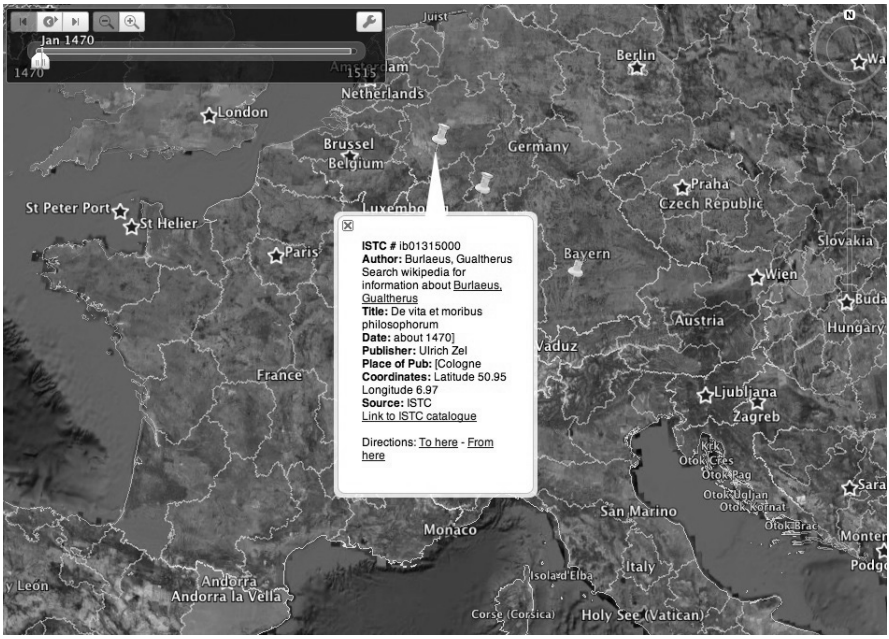

Fig. 1. Geographic Searching in the Incunabula Short Title Catalog (ISTC)

their own set of challenges for efficient implementation. Among these challenges is the fact that the most popular interface for search systems (the simple search box), is extremely cumbersome for entering geographic searches based on coordinates. Users will seldom, if ever, know accurate coordinates for the places they are interested in. They can, however, often find them on a map. In this demonstration we will show how map-based interfaces (using Google Earth and Google Maps) can be used in conjunction with GIR search methods for retrieval of digital library information.

1.1 Probabilistic Spatial Ranking

A search method that employs the "Probability Ranking Principle", is one in which information objects are ranked and presented to the user in decreasing order of their estimated probability of relevance to the user's information need[6]. In previous work [5,1] we have described the development and testing of a probabilistic GIR retrieval model based on logistic regression. The form of that model used in this demonstration estimates the probability of relevance for a particular query and particular record in the database $P(R \mid Q, D)$, using the equivalent "log odds" of relevance expressed $\log O(R \mid Q, D)$ for a set of coefficients, c_i, associated with a set of S statistics, X_i, derived from the query and database, such that:

$$\log O(R \mid Q, D) = c_0 \sum_{i=1}^{S} c_i X_i \tag{1}$$

Fig. 2. Geographic Searching in the Congressional Biography Database

where c_0 is the intercept term of the regression. The spatial ranking, or probability of relevance, can then be simply determined from the log odds. For our retrieval approach, the explanatory statistics or feature variables of Geographic Information Objects (GIOs, i.e., the georeferenced items in the database being searched) included in the logistic regression model are fairly simple:

$X_1 =$ area of overlap(query region, candidate GIO) / area of query region
$X_2 =$ area of overlap(query region, candidate GIO) / area of candidate GIO

X_1 and X_2 are based on the extent of the area of overlap and non-overlap between the query and candidate regions. As described in [1] the c_i coefficients were estimated from a sample of geographic documents, and the resulting algorithm was tested on a different experimental set, showing significantly better performance than any previously described geographic ranking algorithm. In addition we will show how text search can be effectively used in mixed geographic and topical search context using another logistic regression-based algorithm based on our results from the GeoCLEF evaluation[4].

In the search system that will be demonstrated, we use the user's interaction with Google Earth to determine the query region, i.e., the query is based on the user's current view of the world as seen in Google Earth, specifically the bounding coordinates of the area currently visible. A new query is sent to the

search system each time the user changes their view by moving, zooming in, or zooming out. The algorithm above is used to search for data in the database that overlap that search region using the algorithm described above (in the case of point data, the candidate GIO is assumed to be a small region surrounding the geographic point. Figures 1 and 2 show screen shots of this interface for data from the British Library's Incunabula Short Title Catalog (ISTC), and from an RDF database of events in the lives of U.S. Congressional representatives and Senators. The individual georeferenced items in these collections are automatically linked to other topically related databases including the official ISTC database site, Wikipedia, and the official US Congressional Biography site.

Acknowledgments

The work presented draws on two projects partially supported by Institute of Museum and Library Services National Leadership Grants: Support for the Learner: What, Where, When, and Who (http://ecai.org/imls2004/) and Bringing Lives to Light: Biography in Context (http://ecai.org/imls2006/). The system, and demo, would not have been possible without the work of PhD students Patricia Frontiera and Ryan Shaw and the efforts of Co-PIs Michael Buckland and Fredric C. Gey.

References

1. Frontiera, P., Larson, R.R., Radke, J.: A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science 22(3), 337–360 (2008)
2. Hill, L.L.: Georeferencing: The Geographic Associations of Information. MIT Press, Cambridge (2006)
3. Larson, R.R.: Geographic information retrieval and spatial browsing. In: Smith, L., Gluck, M. (eds.) GIS and Libraries: Patrons, Maps and Spatial Information, pp. 81–124. University of Illinois at Urbana-Champaign, GSLIS, Urbana-Champaign (1996)
4. Larson, R.R.: Cheshire at GeoCLEF 2008: Text and fusion approaches for GIR: CLEF working notes (2008), `http://www.clef-campaign.org/2008/working_notes/larson_GeoCLEF.pdf`
5. Larson, R.R., Frontiera, P.: Spatial ranking methods for geographic information retrieval (gir). In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 45–56. Springer, Heidelberg (2004)
6. Robertson, S.E.: The probability ranking principle in ir. Journal of Documentation 33, 294–304 (1977)