# Ranking and Representation for Geographic Information Retrieval

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

## 1. EXTENDED ABSTRACT

Unlike related work in GIS and spatial analysis, there has been little study of the use and effectiveness of different spatial approximations for GIR. Clearly, no one approach will apply to all types of geographic information resources or digital libraries. The need to retrieve and evaluate information objects based on their geospatial characteristics increases as the geographical "aboutness" of the objects increase, e.g. a guidebook vs. a digital geographic data set of invasive plant species. Moreover, geospatial ranking methods are becoming increasingly important as the supply of and demand for geographic information grows. The quality of geospatial approximations in GIR, i.e. how closely they represent the original objects, constrains how accurately and effectively these objects can be retrieved and ranked[8].

We have been exploring these issues and have developed some new algorithms for ranked retrieval of georeferenced objects. We have also been examining the indexing methods that can be employed for materials with geographic content or associations. We have been doing a comparative analysis of several GIR algorithms and evaluating their relative performance using a test collection of geospatial metadata, derived from the California Environmental Information Catalog (CEIC – http://ceres.ca.gov/catalog). This discussion is based on that work.

### 1.1 Geospatial Metadata

Geographic digital libraries typically use geospatial metadata to provide surrogate representations of geographic resources that encode the structure and content of digital geographic data to support identification, discovery, evaluation, and understanding. This metadata is vital for most geographic data because, as non-textual, abstract representations of complex phenomena, they cannot be effectively and appropriately used without it.

In this study we used metadata based on the FGDC Standard, which was created specifically to describe digital geospatial data, but which has also been applied to paper maps, air photos, atlases, environmental impact statements, and other geographically related materials. The elements in this data that concern the geospatial characteristics of the data include: *Spatial domain* (geographic coordinates defining the data's extent), *Place names* (qualitative descriptors of the geographic extent), *Spatial reference system* (projection and coordinate system information, *Spatial Representation model* (vector, raster), *Spatial features* (type and quantity) and *Spatial data quality* (accuracy, completeness, lineage, and sources).

Copyright is held by the author/owner.
*SIGIR '04* July 25-28, 2004, Sheffield, U.K.
ACM 0-89791-88-6/97/05.

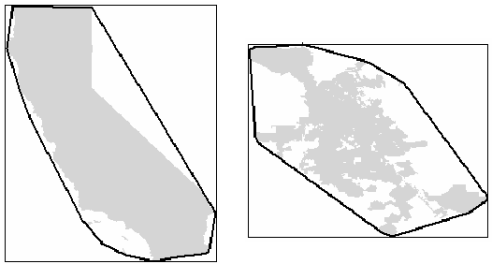

**Figure 1: MBRs (thin lines) and Convex Hulls for the State of California and the City of San Jose.**

However, the only geospatial element required by the FGDC Standard is Spatial Domain. While this element permits complex spatial representations, only a coordinate pair that defines a minimum bounding rectangle (MBR) of the object is required.

As can be seen in Figure 1, MBRs provide a compressed, abstract approximation of a spatial object. The representation is conceptually powerful because it evokes a printed map. Its simplicity, computational efficiency, and storage advantages make it the most commonly used spatial approximation[3]. Yet, the MBR has obvious weaknesses when representing diagonal, irregular, non-convex, or multi-part regions[11]. MBRs over-estimate area, misrepresent shape, and fail to capture the distribution of the data within themselves, leading to "false positives" in GIR matching.

Other spatial approximations, such as the minimum bounding ellipse, minimum bounding N-corner convex polygon, and convex hull, have been investigated in the context of spatial databases and GIS applications, but not for GIR, where the MBR still represents the state of the art. In searching, a query region representing the user's area of interest may be defined by 1) Entering geographic coordinates for a point or bounding box, 2) Using a graphical map interface to zoom in to, click on, or draw a polygon, typically a bounding box, around the area of interest and 3) Entering a place name or selecting it from a list.

The first two methods result in the delineation of a coordinate-based query region. The third uses a digital gazetteer to ob-

| Reference | Formula |
|---|---|
| Hill, 1990[6] | $Range = 2\frac{O}{Q+C}$ |
| Walker et al, 1992[12] | $Range = MIN\left(\frac{O}{Q}, \frac{O}{C}\right)$ |
| Beard and Sharma, 1997[2] | Case 1: Q contains C<br>$Range = \frac{C}{Q}$<br>Case 2: Q and C overlap<br>$Range = \frac{O/Q\%}{(1-O/C)\%+100}$<br>Case 3: Q contained in C<br>$Range = \frac{Q}{C}$ |
| Where:<br>$Q$ = area of query region<br>$C$ = area of candidate GIO<br>$O$ = area of overlap for $G, C$ | Range (for all):<br>0 = no similarity<br>1 = identical |

**Table 1: Methods for computing spatial similarity.**

tain coordinate representations for named places. Regardless of the method used, a query region is often represented internally as a simple bounding rectangle[7]. For geospatial searches, the query region is compared with MBRs of all candidate geographic information objects (GIOs) in the digital library using polygon-polygon geometric operations. If there is overlap between the query and the GIO regions, the GIO is considered a match.

## 1.2 Similarity Measures and Spatial Ranking

GIR ranking methods are based on quantifying the similarity between the query and a GIO in the collection. This similarity "score" can be interpreted as an estimate of the relevance, or utility, of a candidate GIO for a user's information need. There are three basic approaches to spatial similarity measures and ranking:

**Method 1: Simple Overlap.** Candidate geographic information objects (or GIOs) that have any overlap with the query region are retrieved.

**Method 2: Topological Overlap.** Spatial searches are constrained to only those candidate GIOs that: a) are completely contained within, b) overlap, or c) contain the query region. Each category is exclusive and all retrieved items are considered relevant.

**Method 3: Extent of Overlap.** A spatial similarity score is derived from the extent of overlap between a candidate GIO and the query region. The greater the overlap, the greater the assumed relevance of the candidate GIO to the query. A variety of spatial scores based on overlap are discussed in the literature (Hill, 1990; Walker et al, 1992; Beard and Sharma, 1997) and presented in Table 1.

The simple and topological overlap approaches are most commonly used in digital libraries where the geographic objects of interest are represented by MBRs. Retrieval algorithms based on MBRs are easy to implement and are supported by the GEO profile of the Z39.50 information retrieval protocol[10]. However, the Boolean matching criterion does not allow for spatial ranking and thus inhibits good retrieval performance [1](p. 26), especially as result sets grow in size. Classifying retrieved candidates based on topological relationships (e.g., contains, overlaps, contained within), as in method 2, is a first step in discriminating among the results, but it doesn't speak directly to the issue of relevance. Moreover, the burden is on the user to understand these relationships and how they impact a geospatial search. There has been very limited research on the effectiveness of spatial ranking with Hill[6] presenting the only empirical data and evaluation.

Clearly, many research questions concerning spatial ranking need to be investigated, including ways in which it can be implemented and evaluated. These questions become increasing critical as the amount of geographic information in digital libraries, and thus the size of result sets, continues to grow.

## 1.3 Our Approach: Probabilistic Spatial Ranking

Maron and Kuhns[9] first introduced the idea that, given the imprecise and incomplete ways in which a user's information need is represented by a query and an information object by its indexing, relevance should be approached probabilistically. This is especially true for geographic information retrieval since all geographic information objects are abstract, compressed representations of real world phenomena that contain some degree of error and uncertainty[5].

In the logistic regression (LR) model of IR[4], the estimated probability of relevance for a particular query and a particular record in the database $P(R \mid Q, D)$ is calculated as the "log odds" of relevance $\log O(R \mid Q, D)$ and converted from odds to a probability. The LR model provides estimates for a set of coefficients, $c_i$, associated with a set of $S$ statistics, $X_i$, derived from the query and database, such that:

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

where $c_0$ is the intercept term of the regression. The spatial ranking, or probability of relevance, can then be given as:

$$P(R \mid Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \quad (2)$$

For this study, the geospatial characterics, i.e. explanatory statistics or feature variables, explored in the logistic regression model are:

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

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

$X_3$ = 1 - abs(fraction of query region that is onshore - fraction of candidate GIO that is onshore)

Like the spatial similarity measures presented in Table 1, $X_1$ and $X_2$ are based on the extent of the area of overlap and non-overlap between the query and candidate GIO regions. $X_3$ requires a bit more explanation. As noted in Hill[6] geographic areas that are near a coastline can be problematic when approximated by simplified geometries like the MBR. The MBR for an offshore region may necessarily include a lot of onshore area, and vice versa. We define $X_3$ as a "shorefactor" variable that captures the similarity between the fraction of a query region that is onshore compared to that of a candidate GIO region. For example, if a query region is 20% onshore and a candidate GIO region is 75% on

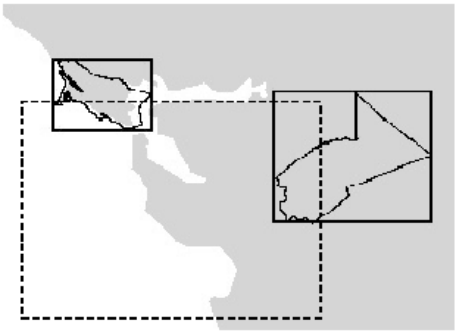

**Figure 2: Search Query (dashed rectangle) and MBRs and Polygon Representations of Marin (NW) and Stanislaus (E) Counties.**

shore, then the shorefactor is $1 - abs(.20 - .75) = .45$. Calculating shorefactor is illustrated in Figure 2. Marin County is 70% onshore, while Stanislaus County is 100% onshore. The dashed query box in Figure 2 is 45% onshore. Thus, the shorefactor for Marin is $1 - abs(.45 - .70) = .75$ while for Stanislaus it is $1 - abs(.45 - 1) = .45$. A shorefactor of 1 indicates that both regions are either offshore or onshore. A shorefactor approaching 0 indicates that one region is almost completely onshore and one is almost completely offshore, thus it allows geographic context to be integrated into the spatial ranking process.

The shorefactor was computed by intersecting both the query and GIO regions with a very generalized polygonal representation of the Western USA.

The results of our analysis of these algorithms, for both MBRs and Convex hulls has been submitted to the European Conference on Digital Libraries and will be describe at the Workshop.

## 2. REFERENCES


[1] R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*. Addison Wesley, New York, 1999.

[2] K. Beard and V. Sharma. Multidimensional ranking for data in digital spatial libraries. *International Journal of Digital Libraries*, 1(2):153–160, 1997.

[3] T. Brinkhoff, H. P. Kriegel, and R. Schneider. Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems. In *Proceedings of 9th International Conference on Data Engineering*, 1993.

[4] W. S. Cooper, F. C. Gey, and D. P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.

[5] M. Goodchild. Future directions in geographic information science. *Geographic Information Science*, 5(1):1–8, 1999.

[6] L. L. Hill. *Access to Geographic Concepts in Online Bibliographic Files: effectiveness of current practices and the potential of a graphic interface*. PhD thesis, University of Pittsburgh, Pittsburgh, 1990.

[7] L. L. Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha and T. Baker, editors, *Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 (Lisbon, Portugal, September 18-20, 2000)*, pages 280–290, Berlin, 2000. Springer.

[8] C. B. Jones, H. Alani, and D. Tudhope. *Geographical Terminology Servers – Closing the Semantic Divide*, chapter 11, pages 205–222. Taylor and Francis, London, 2003.

[9] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.

[10] D. D. Nebert. Z39.50 application profile for geospatial metadata or 'GEO', version 2.2, 27 may 2000. Available as: http://www.blueangeltech.com/Standards/GeoProfile/geo22.htm.

[11] D. Papadias, Y. Theodoridis, T. Sellis, and M. Egenhofer. Topological relations in the world of minimum bounding rectangles: a study with r-trees. In *Proceedings of the ACM SIGMOD Conference, San Jose, California*, 1995.

[12] D. Walker, I. Newman, D. Medyckyj-Scott, and C. Ruggles. A system for identifying datasets for gis users. *International Journal of Geographical Information Systems*, 6(6):511–527, 1992.