Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries*

Ray R. Larson[1] and Patricia Frontiera[2]

[1] School of Information Management and Systems
ray@sherlock.berkeley.edu
[2] College of Environmental Design
University of California, Berkeley
Berkeley, California, USA, 94720
pattyf@regis.berkeley.edu

Abstract. This paper presents results from an evaluation of algorithms for ranking results by probability of relevance for Geographic Information Retrieval (GIR) applications. We review the work done on GIR and especially on ranking algorithms for GIR. We evaluate these algorithms using a test collection of 2500 metadata records from a geographic digital library. We present an algorithm for GIR ranking based on logistic regression from samples of the test collection. We also examine the effects of different representations of the geographic regions being searched, including minimum bounding rectangles, and convex hulls.

1 Introduction

Geographic data are an extremely important resource for a wide range of scientists, planners, policy makers, and analysts who study natural and planned environments. Notably, the landscape of geographic analysis has been changing rapidly from data and computation poor to data and computation rich[15]. Developments in digital electronic technologies, such as satellites, integrated GPS units, digital cameras, and miniature sensors, are dramatically increasing the types and amounts of digitally available raw geographic data and derived information products[17]. At the same time, advances in computer hardware, software and network technologies continue to improve our ability to store and analyze these large, complex data sets.

These factors are contributing to a growing political, social, scientific and economic awareness of the value of geographic information and driving new applications for its use. In response to this, geographic digital libraries are growing in number, collection size, and sophistication. Moreover, mainstream digital libraries, i.e. those that deal with primarily text materials, are increasingly considering geographic access methods for information resources that have important

* *This is a preprint of a paper accepted for the 2004 European Conference on Digital Libraries in Bath, U.K. The proceedings will be published by Springer-Verlag in the Lecture Notes in Computer Science Series.*

geographic characteristics. 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.

One common approach in digital libraries is to use place names as a geographical search surrogate. However, place names have well-documented lexical and geographical problems [13]. 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. Geographic coordinates, on the other hand, provide an unambiguous and persistent method for locating geographic areas or features. However, the use of coordinates presents many challenges in terms of storage, indexing, processing and user interface design that only recently have begun to be investigated in the context of geographic information retrieval (GIR) for digital libraries.

One key question for GIR is what level of detail should be used to encode coordinate information? Gazetteer research cautions that complex spatial objects present enormous data storage and performance problems for online geographic digital libraries[1, 11], which provide, at best, extremely limited GIS functionality. The decomposition of complex spatial objects into approximate representations is a common approach to simplifying coordinate representations. Early work in this area by Hill [10] suggests that minimum bounding rectangles can sufficiently represent geographic objects for information retrieval applications. Other research[1, 12] indicates that even single point representations can be used effectively when combined with innovative retrieval and ranking methods.

In this paper we explore these issues and present some new algorithms for ranked retrieval of georeferenced objects in digital library collections. We discuss the characteristics of georeferenced information and its use in digital libraries. The next section describes the primary components for GIR within digital libraries and describes the characteristics of GIR in a digital library context. Subsequent sections examine indexing and access creation for geo-referenced sources. We then examine the retrieval effectiveness of several GIR algorithms using a test collection of geospatial metadata from the California Environmental Information Catalog (CEIC – http://ceres.ca.gov/catalog).

2 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.

Geospatial metadata specifically addresses the encoding of coordinate representations of geographic objects. There are geospatial metadata standards in most EU countries. In the U.S., it is usually created in accordance with one of two metadata standards: 1) the Dublin Core (DC)[6]; or 2) the Federal Geo-

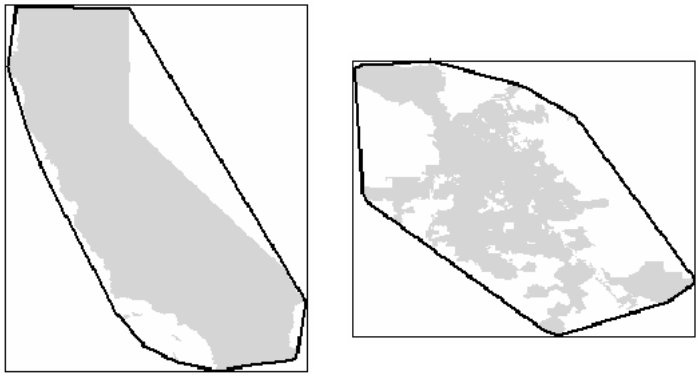

Fig. 1. MBRs (thin lines) and Convex Hulls for the State of California and the City of San Jose.

graphic Data Committee's Content Standard for Digital Geospatial Metadata (FGDC)[8]. The only geographic element in the base DC is the Coverage element. This element can be used to specify a place name, place code (e.g. zip code), or the geospatial coordinates of a point, bounding rectangle, or irregular polygon that locates the resource being described.

The FGDC Standard was created specifically to describe digital geospatial data, but is also applied to paper maps, air photos, atlases, environmental impact statements, and other geographically related materials. It provides elements that address the geospatial characteristics of the data, including: *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). The FGDC Standard *requires* only a coordinate pair defining a minimum bounding rectangle (MBR) for the object, but allows more complex descriptions also.

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[4]. Yet, the MBR has obvious weaknesses when representing diagonal, irregular, non-convex, or multi-part regions[18]. MBRs over-estimate area, misrepresent shape, and fail to capture the distribution of the data within themselves, leading to "false positives" in GIR matching.

3 Geospatial Search and Ranking Methods

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

Reference	Formula
Hill, 1990[10]	$Range = 2\frac{O}{Q+C}$
Walker et al, 1992[19]	$Range = MIN\left(\frac{O}{Q}, \frac{O}{C}\right)$
Beard and Sharma, 1997[3]	Case 1: Q contains C $Range = \frac{C}{Q}$ Case 2: Q and C overlap $Range = \frac{O/Q\%}{(1-O/C\%)+100}$ Case 3: Q contained in C $Range = \frac{Q}{C}$
Where: Q = area of query region C = area of candidate GIO O = area of overlap for G, C	Range (for all): 0 = no similarity 1 = identical

Table 1. Methods for computing spatial similarity.

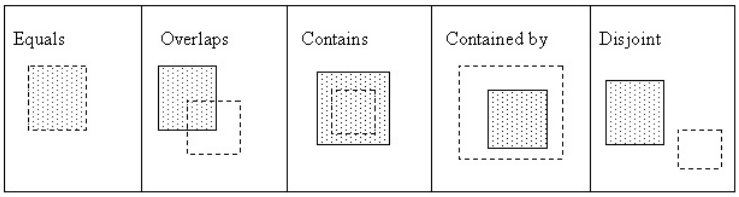

Fig. 2. Spatial relationships between overlapping regions.

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 obtain coordinate representations for named places. Regardless of the method used, a query region is often represented internally as a simple bounding rectangle[11]. 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. Possible relationships between two overlapping regions are illustrated in Figure 2. This is a simplified subset of the 9 intersection topological model for spatial relations[7]. Proximity relationships (such as near or adjacent) are not considered matches.

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. Retrieved items are ranked and presented to the user in descending order of these scores. While traditional IR scores and rankings are based on the statistical properties of terms in a collection, GIR relies on spatial scores and

rankings based on geospatial characteristics such as size, shape, location, and distance. 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[16]. However, the Boolean matching criterion does not allow for spatial ranking and thus inhibits good retrieval performance [2](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[10] presenting the only empirical data and evaluation.

3.1 Probabilistic Spatial Ranking

Maron and Kuhns[14] 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[9].

In the logistic regression (LR) model of IR[5], 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:

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

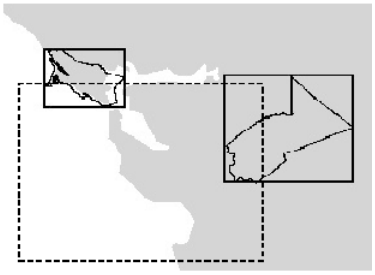

Fig. 3. Search Query (dashed rectangle) and MBRs and Polygon Representations of Marin (NW) and Stanislaus (E) Counties.

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

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

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[10] 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 shore, then the shorefactor is $1 - abs(.20 - .75) = .45$. Calculating shorefactor is illustrated in Figure 3. Marin County is 70% onshore, while Stanislaus County is 100% onshore. The dashed query box in Figure 3 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.

4 Evaluation Approach

We applied our logistic regression method and the three spatial ranking methods presented in Table 1 to a test collection of geospatial metadata. The research questions were: 1) How effectively can the geographic relationship between a query region and the region of a candidate information object be evaluated and ranked based on the overlap of the geographic approximations of these regions? and 2) How do different geographic approximations affect the rankings? To examine question one, the results of the different ranking methods were summarized and compared using two standard retrieval performance evaluation measures: average precision at 11 standard recall levels and the mean average query precision.

In response to the second question, we applied and compared the results of these ranking methods for both MBR and convex hull approximations of the candidate GIOs. The convex hull is the minimum convex polygon that contains a geometric object (i.e. collection of points). It can be visualized as a rubber band around a geographic polygon to approximate its extent (see heavy lines in Figure 1). The convex hull is widely used as a geometric approximation in GIS and it provides the best approximation quality of conservative (i.e., encloses all points of the original) convex representations[4]. Because a convex hull is a better approximation of the original spatial object than an MBR, it will retrieve fewer false positives when used for GIR.

We assumed that candidate GIO regions that overlap the query region are relevant and regions that do not overlap are not relevant. Given that all regions are represented by conservative approximations, all relevant items will be retrieved (i.e. 100% recall). However, not all *approximations* that overlap will be relevant because the regions they represent may *not* overlap[18].

4.1 Test Collection Overview

The test collection for this study was a subset of metadata records from the California Environmental Information Catalog (CEIC), (http://ceres.ca.gov/catalog). The CEIC collection includes a wide variety of different types of geographic information resources, including: vector and raster geospatial data, maps, databases, documents, reports, websites, models, etc. These resources are documented with metadata prepared in accordance with the FGDC standard. For this study, approximately 2500 metadata records in XML format were selected from the total collection of about 4000 (as of August 2003). These records can be divided into two main categories: 1) those that refer to known, named geographic regions within the state; and 2) user defined areas (UDAs) - those regions that are specific to the person or organization that created the GIO described by the metadata. An important distinction between these two categories is that the geographic regions associated with the CA places are typical of those found in gazetteers and place name thesauri. Moreover, these regions can be traced, via their names, to geographic data containing more precise geographic representations, which we used in calculating the "shorefactor" described above. For the

UDA regions, which seldom have accurate or complete data available, we assume that the both the convex hull and complex polygon representations of the geographic extent are equal to the MBR approximation. The MBRs and convex hulls were pre-processed in the ESRI ArcView software and then loaded into Postgres 7.4, with the PostGIS 0.8 and GEOS 1.0 extensions, where the analysis was done.

5 Results

This research considers the test collection in two parts. In the first part, the issues of spatial representation and ranking are considered for the metadata indexed by CA places. The second part considers the entire collection, both CA places and UDAs.

Our first set of tests considered only the 2072 test metadata records, or GIOs, that were indexed geospatially by known CA places. The 42 CA counties referenced in these GIOs were used, each in turn, as query regions. MBR and convex hull approximations of all CA places referenced by these metadata were treated as candidate GIO regions.

The first task was to determine the reference set of candidate GIO regions relevant to each county query region. This was done using the complex polygon data to select all CA place regions that overlap, contain, or are contained within the query region. All retrieved regions were reviewed (semi-automatically) to remove sliver matches, i.e. those regions that only overlap due to inconsistencies in the data. This process resulted in a master file of CA place regions relevant to the 42 CA county query regions. Queries for ten county regions were used to train the logistic regression models. LR Equation 3 was used for the MBR rankings and LR Equation 4 for convex hulls:

$$\log O(R \mid Q, D) = -5.040 + (6.5154 \cdot X_1) + (5.7729 \cdot X_2) \tag{3}$$

$$\log O(R \mid Q, D) = -3.4767 + (7.4536 \cdot X_1) + (5.7569 * X_2) \tag{4}$$

Queries for the other 32 county query regions were run against the MBR and convex hull approximations of the candidate GIO regions. We then applied all four spatial ranking methods to the result sets and calculated precision-recall summary statistics.

Tables 2 and 3 show the evaluation results of the four spatial ranking methods on the CA places subset of the test collection. These tables show that: 1) the values for the non-logistic regression ranking methods are extremely similar; 2) the logistic regression method performed better than the other methods on this test collection; 3) for all methods, rankings based on the convex hull representations performed better than those based on the MBR representations. Yet, it is interesting to note that the non-logistic regression spatial ranking methods applied to the convex hull approximations do *not* perform better than the logistic regression method applied to the MBRs. An important implication of this is that it is worth investigating more effective spatial ranking methods before adopting

Ranking method	MBRs	Convex Hulls
Hill, 1990	0.7193	0.8097
Walker, et al., 1992	0.7025	0.8006
Beard and Sharma. 1997	0.7094	0.8116
Logistic Regression	0.9389	0.9973

Table 2. Mean Average Query Precision for Named Places.

	Minimum Bounding Rect.				Convex Hulls			
Recall Level	Hill	Walker	Beard	Logistic	Hill	Walker	Beard	Logistic
0.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.10	0.8668	0.8660	0.8717	0.9777	0.9232	0.9248	0.9277	1.0000
0.20	0.8409	0.8362	0.8430	0.9663	0.9152	0.9049	0.9083	1.0000
0.30	0.8101	0.8109	0.8214	0.9651	0.8708	0.8775	0.8813	1.0000
0.40	0.8017	0.7985	0.8073	0.9651	0.8705	0.8669	0.8746	1.0000
0.50	0.7940	0.7972	0.8068	0.9651	0.8661	0.8658	0.8735	1.0000
0.60	0.7919	0.7951	0.8039	0.9651	0.8660	0.8658	0.8735	1.0000
0.70	0.7919	0.7951	0.8039	0.9643	0.8623	0.8658	0.8698	0.9997
0.80	0.7919	0.7951	0.8039	0.9520	0.8623	0.8658	0.8698	0.9983
0.90	0.7914	0.7947	0.8035	0.9477	0.8621	0.8656	0.8696	0.9882
1.00	0.7613	0.7684	0.7881	0.9114	0.8243	0.8274	0.8291	0.9648
Avg Prec	0.8220	0.8234	0.8321	0.9618	0.8839	0.8846	0.8888	0.9955

Table 3. Average Precision at 11 Standard Recall Levels for Named Places using Minimum Bounding Rectangles and Convex Hulls.

more complex spatial approximations. Average precision at 11 standard recall levels (Table 3) gives one an idea of how an algorithm performs over the course of retrieving all relevant GIOs. Mean average query precision is a measure that favors systems that rank relevant documents early in the results (Table 2). It averages precision values after each new relevant document is observed in the ranked list and presents a summary statistic of overall performance. However, it may not indicate if an algorithm has poor recall [2](p. 80). But, these characteristics make it a good fit for spatial ranking algorithms because poor recall is very rarely an issue and high precision is desireable. Moreover, the metric is insensitive to differences in the number of items indexed to the same geographic region. This latter is not true of average precision at standard recall levels, therefore the values for average precision (Table 3) are lower than those for mean average query precision (Table 2). Interestingly, the difference between these values for the logistic regression method does not differ as much as for the non-logistic regression ranking methods.

The second part of our study considers the test collection metadata as a whole: both those metadata indexed by CA places and those indexed by user-defined areas (UDAs). As with the tests in Part I, the 42 CA counties referenced in the GIOs were considered query regions and the MBR and convex hull representations for all geospatially indexed areas were treated as candidate GIO regions.

Ranking Method	MBRs	Convex Hulls
Ranking Method	MBRs	Convex Hulls
Hill, 1990	0.6722	0.7936
Walker et al., 1992	0.6509	0.7810
Beard and Sharma, 1997	0.6523	0.7778
Logistic Regression 1	0.8141	0.9099
Logistic Regression 2	0.8819	0.9238

Table 4. Mean Average Query Precision for Full Collection

The reference set of UDA regions relevant to each county query region was determined through a manual review of the UDA metadata. This process could not be automated because, unlike the CA place regions, there are no reference data sets of complex polygons that delineate the UDA regions.

As in Part I, queries for ten county query regions were used to train the logistic regression models. Because 88% of the UDAs represent coastal or offshore regions, an additional logistic regression model was tested that includes the shorefactor variable. LR equations 5 and 6 were used for MBRs and equations 7 and 8 were used for convex hulls.:

$$\log O(R \mid Q, D) = -1.6747 + (1.9871 \cdot X_1) + (3.2976 \cdot X_2) \quad (5)$$

$$\log O(R \mid Q, D) = -2.1303 + (1.9138 \cdot X_1) + (3.2157 \cdot X_2) + (0.7451 \cdot X_3) \quad (6)$$

$$\log O(R \mid Q, D) = -1.2123 + (1.4471 \cdot X_1) + (5.4585 \cdot X_2) \quad (7)$$

$$\log O(R \mid Q, D) = -1.2825 + (1.4341 \cdot X_1) + (5.4096 \cdot X_2) + (0.1267 \cdot X_3) \quad (8)$$

The evaluation results are presented in Tables 4 and 5. These show a similar pattern to the results presented in Part I. The logistic regression rankings perform better than non-logistic regression methods and the convex hull approximations also perform better than the MBRs. Again, the logistic regression rankings for MBRs perform as well as or better than the non-logistic regression rankings for convex hulls, although by a smaller margin than when just the CA place regions were considered.

The addition of the UDA regions significantly degrades the retrieval performance for all algorithms, even though these regions only index 19% of the total metadata records. The majority of the UDA regions are for coastal or near-coastal offshore areas which, when approximated by either MBRs or convex hulls necessarily overlap with onshore regions, thus generating more false-positive retrievals. The logistic regression model (LR2) that incorporates the shorefactor variable is meant to address this problem, yet this method shows only a small (but significant) improvement over the other logistic regression model (LR1), especially for the convex hull approximations. T-tests for paired samples of LR1 and LR2 results gave results ranging from -3.028 to -4.144 with 0.005 or less probability of random occurrence.

	Minimum Bounding Rect.					Convex Hulls				
Recall Level	**Hill**	**Walker**	**Beard**	**LR1**	**LR2**	**Hill**	**Walker**	**Beard**	**LR1**	**LR2**
0.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.10	0.8384	0.8453	0.8562	0.9174	0.9529	0.9099	0.9146	0.9188	0.9715	0.9781
0.20	0.7983	0.7782	0.7885	0.9114	0.9413	0.8905	0.8827	0.8874	0.9676	0.9746
0.30	0.7575	0.7729	0.7871	0.8898	0.9395	0.8484	0.8634	0.8686	0.9560	0.9641
0.40	0.7402	0.7460	0.7570	0.8785	0.9310	0.8428	0.8482	0.8585	0.9534	0.9635
0.50	0.7377	0.7450	0.7570	0.8767	0.9291	0.8406	0.8481	0.8583	0.9534	0.9625
0.60	0.7350	0.7420	0.7538	0.8742	0.9291	0.8406	0.8481	0.8583	0.9534	0.9625
0.70	0.7350	0.7420	0.7538	0.8742	0.9291	0.8403	0.8481	0.8548	0.9505	0.9579
0.80	0.7350	0.7420	0.7538	0.8631	0.9182	0.8371	0.8481	0.8548	0.9412	0.9539
0.90	0.7344	0.7416	0.7534	0.8631	0.9018	0.8312	0.8478	0.8544	0.9342	0.9432
1.00	0.7067	0.7139	0.7311	0.7715	0.7743	0.7819	0.7782	0.7787	0.8340	0.8272
Avg Prec	0.7744	0.7790	0.7902	0.8836	0.9224	0.8603	0.8661	0.8721	0.9468	0.9534

Table 5. Average Precision at 11 Standard Recall Levels for the Full Collection using Minimum Bounding Rectangles and Convex Hulls

6 Conclusions

In GIS and spatial database technologies, geometric approximations, primarily the MBR, are used as a first step to filter possible matches. Then, a refinement step examines the actual complex spatial objects to determine the final result set. However, in a geographic digital library environment, the end-user is the refinement step. For this reason, both high-quality approximations that limit the number of false matches and spatial ranking strategies that present best matches first are extremely important in GIR. We have shown that a logistic regression based spatial ranking algorithm can provide significant improvements for geographic information retrieval, even when the simplest regional approximations (MBRs) are used. We have also shown that taking into account the portion of offshore areas included in a geographic representation can improve GIR performance even further.

7 Acknowledgments

Work on geographic information retrieval 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".

References

1. H. Alani, C. B. Jones, and D. Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. *Internation Journal of Geographical Information Science*, 15(4):287–306, 2001.

2. R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval*. Addison Wesley, New York, 1999.
3. K. Beard and V. Sharma. Multidimensional ranking for data in digital spatial libraries. *International Journal of Digital Libraries*, 1(2):153–160, 1997.
4. 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.
5. 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.
6. Dublin Core Metadata Initiative. Dublin core metadata element set, version 1.1: Reference description, 2003-02-04. Available as: http://dublincore.org/documents/2003/02/04/dces/.
7. M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991.
8. Federal Geographic Data Committee. Content standard for digital geospatial metadata,csdgm version 2 - fgdc-std-001-1998. Available as: http://www.fgdc.gov/standards/documents/standards/metadata/v2_0698.pdf.
9. M. Goodchild. Future directions in geographic information science. *Geographic Information Science*, 5(1):1–8, 1999.
10. 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.
11. 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.
12. 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.
13. R. R. Larson. Geographic information retrieval and spatial browsing. In L. Smith and M. Gluck, editors, *GIS and Libraries: Patrons, Maps and Spatial Information*, pages 81–124. University of Illinois at Urbana-Champaign, GSLIS, Urbana-Champaign, 1996.
14. M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.
15. H. J. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery: An Overview*. Taylor & Francis, New York, 2001.
16. 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.
17. S. Openshaw. Geographical data mining: key design issues. In *GeoComputation'99*, 1999.
18. 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.
19. 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.