A Spatial Overlay Ranking Method for a Geospatial Search of Text Objects

By Kenneth J. Lanfear, USGS Retired

Open-File Report 2006-1279

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
Dirk A. Kempthorne, Secretary

U.S. Geological Survey
P. Patrick Leahy, Acting Director

U.S. Geological Survey, Reston, Virginia 2006

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth,
its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

A Spatial Overlay Ranking Method for a Geospatial Search of Text Objects

Kenneth J. Lanfear, USGS Retired

Abstract

Earth-science researchers need the capability to find relevant information by location and topic. Conventional geographic techniques that simply check whether polygons intersect can efficiently achieve a high recall on location, but can not achieve precision for ranking results in likely order of importance to the reader. A spatial overlay ranking based upon how well an object's footprint matches the search area provides a more effective way to spatially search a collection of reports, and avoids many of the problems associated with an "in/out" (True/False) boolean search. Moreover, spatial overlay ranking appears to work well even when spatial extent is defined only by a simple bounding box.

Background

Earth-science researchers need the capability to find relevant information by location and topic. Modern geographic information systems (GIS's) can identify location matches by quickly finding all points, lines, or polygons that intersect a given area. Ranking the results in likely order of importance to the reader is, however, much more difficult. We want to avoid results like, "Found 10,001 matches to your query. Here they are, listed in no particular order."

Librarians have long dealt with questions of recall and precision. Recall measures how well a search finds all specified objects in a collection. Precision refers to how well only relevant objects are selected or to how well objects are ranked in relevant order. The simplest geospatial search method, the "in/out" Boolean search used by nearly all GIS software, merely looks for objects that intersect any part of the query area. The Boolean search typically executes quickly and achieves high recall. Achieving precision is another matter, since a Boolean search alone can't distinguish a "good" spatial fit from a "bad" spatial fit.

The Boolean search is good enough for many spatial data collections because GIS data sets have metadata that identifies their scale or resolution, and this can be used to improve the precision of a search. A suitable map, for example is one that covers the area of interest and has a scale within a specified range. But what is the scale of a book? Books, reports, web pages, and similar text-based information objects have no explicit resolution. If you ask for a report about the geology of Chicago, is a report on the United States good enough? Is it better or worse than one on Illinois?

Books and other objects imply at least a relative resolution by the extent of their subject area, sometimes called their "footprint." For example, a travel guide of the United States might describe Mount Rushmore in a sentence, whereas a guide of South Dakota might devote several pages to the subject. A person choosing among travel guides would logically select a guide with a footprint that most closely matches his or her area of

interest. Everything else being equal, covering an area bigger or smaller than the desired spatial extent should result in a lower ranking. This is the key to improving precision.

To a query, "Show me information on ground water in Virginia," a search tool should rank its findings for suitability regarding theme (ground water) and spatial extent (the bounds of Virginia). A report titled "Ground Water in Virginia" should rank highly since it deals with the requested theme and covers exactly the area wanted. A report titled "Ground Water in the United States" should rank lower because, although the theme matches, it includes information beyond the spatial extent of the search and presumably provides less detail about Virginia. Similarly, a report on "Ground Water in Fairfax County, Virginia" also should rank lower because, although it might contain great detail about a part of Virginia, it does not cover all of Virginia. The problem is not only one of finding matches – all 3 footprints intersect the search area – but of identifying the result most likely to provide the desired scope and detail.

The process should also be adjustable to user preferences and the nature of the material. In the above example, a user should be able to indicate they'd be happy to find a report with details on a sub-area of their search; this user would not penalize the Fairfax County report for covering only part of Virginia. Similarly, a thick National report still may contain great detail at the State level, and its ranking should reflect this.

Similarity to Linguistic Searches

Hill (1990), examined the geographic similarity between pairs of documents using both linguistic comparisons – basically, comparing geographic keywords – and maps. She found "only weak correlations between text-based and spatially-based geographic representations … related to the imprecise nature of words in representing geographic areas and to the lack of predictability of the terminology used to describe a particular area." (Hill, 1990, p. iv) In doing her comparisons, she explored a variety of techniques for quantitatively expressing the geographic similarity of two areas based upon overlapping areas, common boundaries, and the distance between non-overlapping areas.

The Dice coefficient (Dice, 1945), often used for determining the similarity, S_{XY}, of two sets of keywords, X and Y, is comparable to formulas used for determining the similarity of overlapping areas:

$$S_{XY} = 2\,(X \cap Y) / (X+Y) \qquad \text{(Equation 1)}$$

Similarity between keyword sets is computed as their intersection divided by their union, giving a score of 0 for no match, and 1 for a perfect fit. The spatial overlay method that follows uses almost exactly this principle, translated to coordinate geometry.

Ranking the Relevance of a Spatial Match

Given a query polygon Q, figure 1 shows how we score the spatial relevance of target information object **T**. The intersection, **T** ∩ **Q**, is called **X**.

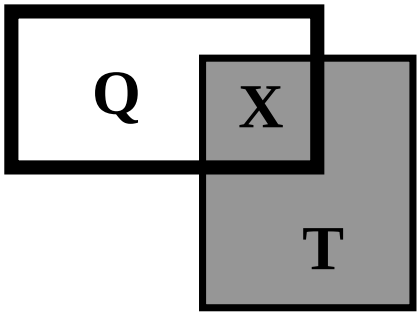

Figure 1. Diagram of query polygon Q intersecting target object T. The intersection area is X.

If the information in **T** is uniformly distributed, then the fraction of the information that is contained in **X** is F_t, where

$F_t = X / T.$ where $0 \le F_t \le 1$ (Equation 2)
and X, T are the areas of **X** and **T**, respectively.

With similar reasoning, we can presume that, since **X** only covers part of **Q**, only a fraction of our information request is fulfilled by **X**, or

$F_q = X / Q$ where $0 \le F_q \le 1$ (Equation 3)
and Q is the area of **Q**.

Using area to compute the fractions assumes a uniform distribution of information within **T** and a uniform importance of information throughout **Q**. Using non-uniform distributions would require more computation, but the principle is the same.

A composite spatial score, S, will be the product of the two fractions,

$S = F_t F_q$. (Equation 4)

The value of S will range from 0 for no match to 1 for a perfect fit.

We can focus the power of this test, increasing the penalty of mismatches, by raising the fractions to powers k_t and k_q for target and query, respectively. The spatial score than becomes,

$S = F_t^{k_t} F_q^{k_q}$ if $F_t, F_q > 0$ (Equation 5)

$S = 0$ otherwise.

Increasing k_t raises the importance of the target being entirely with the query area, and decreasing k_t indicates a willingness to accept targets that extend beyond the query area. Increasing k_q raises the importance of finding targets that cover the whole query area, and decreasing k_q indicates a willingness to accept smaller targets within the query area. Setting either to 0 results in treating that part of the test as a Boolean search.

Some effects of changing the k values can be seen in figure 2. By setting k to a small value, we impose little penalty for a minor mismatch between the target footprint and query footprint. As the upper curve shows, we can set k to control the "break point" where the mismatch penalty quickly mounts. A k value of 0.05, for example, means that a target that overlaps the query area by only 10 percent still scores as high as 0.9.

The spatial score, S, can be multiplied by any thematic score for a combined spatial-thematic score.

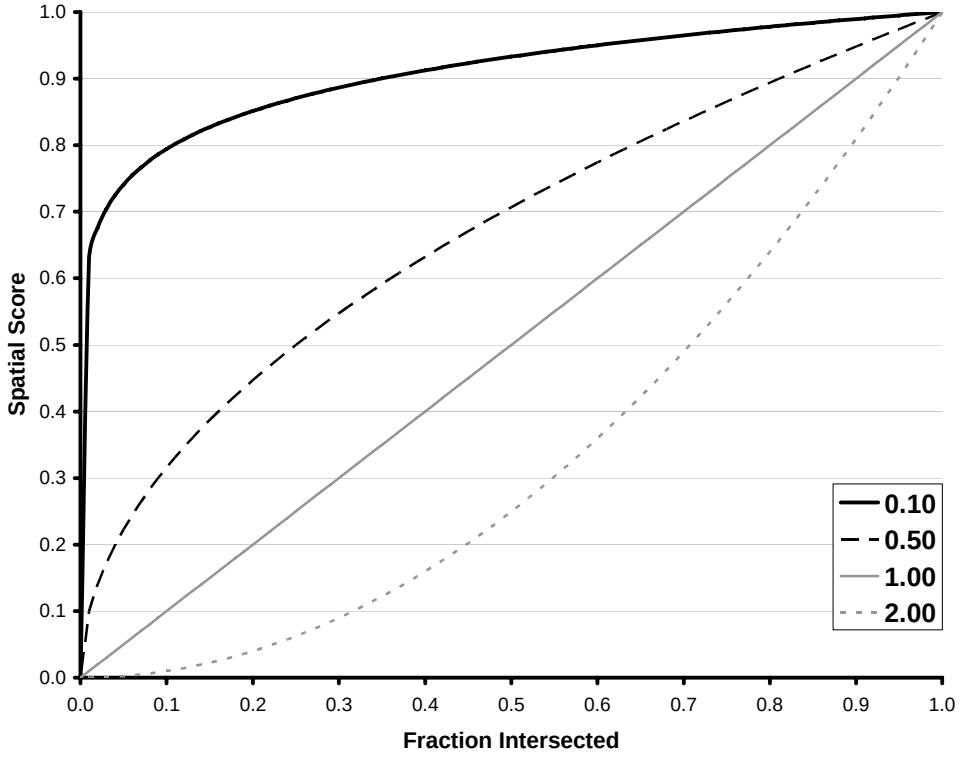

Figure 2. Graph showing how using different K value affects the scores of factional overlays.

Practical Application

The USGS Thesaurus defines a high-level, tree-like hierarchy of categories that describe the science and products of the U.S. Geological Survey (USGS). The *Browse USGS* feature of the USGS website – this feature no longer is available – used to let users navigate around the "tree" of the Thesaurus to explore the categories – their definitions, how they're used, broader or narrower terms, etc. – and find selected USGS web pages that contain more information about a category. Since most USGS information is spatial, the browse also allowed the user to specify an area; it then selected USGS web pages that best fit that area.

Browse USGS selected from among a collection of approximately 2,100 key USGS Web pages (out of more than 400,000 in over 300 USGS Web sites) selected by librarians, webmasters, and subject-matter experts to support the terms of the USGS Thesaurus. These pages were carefully cataloged by category and location. The catalogers determined the footprint of each page as a bounding box of latitude and longitude.

The ability to set the k values in equation 5 turns out to be important. When looking at state- or county-sized areas, a common search criteria, setting $k_t = 0.5$ tends to exclude the big national pages but leaves room for multi-state or regional pages. Setting $k_q = 0.1$ relaxes the requirement for the page to cover the entire query area. Choices are dependent on both the collection characteristics – USGS has many pages on national topics – and on the needs of the user.

As an example of how results are ordered by a spatial overlay ranking, table 1 shows a selection of USGS web pages on volcanic activity in the State of Washington. Note how the selection favors a site specifically about impacts of a volcano in Washington, and gives lesser scores to pages about Washington-Oregon, Washington-Oregon-California, and worldwide, in that order. A Boolean search would have no means of making these distinctions, since Washington is in all of these regions, and would have to present the 19 selections in random or alphabetical order: the best-fitting site would have just over a 50-50 chance of being in the top 10 selections!

Table 1. Example of USGS web pages selected by *Browse USGS* for the category, "volcanic activity" and ranked in descending order of relevance to the spatial extent of Washington. For clarity, importance factors used by *Browse USGS* have been removed.

Web Page	Spatial Extent	Score
Eruptions of Mount St. Helens: past, present, future Online booklet on the 1980 eruption of Mount St. Helens, past history, and future hazards.	Washington	1.0
Cascades Volcano Observatory Portal to information provided by the Cascades Volcano Observatory in Vancouver, WA with links to reports and activity in the Cascade Range and other volcanoes and multiple links to general information on volcanology, reports, research, and maps.	Washington and Oregon	0.68
Volcano video and television surveillance monitoring systems Visual monitoring of volcanoes by closed-circuit television and video monitoring for a permanent video record of events using slow-scan television permits continuous surveillance at a distance for remote volcanoes or in times of danger.	Washington and Oregon	0.68
Volcanoes in Pacific Northwest Online Science Resource Locator	Washington and Oregon	0.68
Cascades Range Volcanoes Weekly Update	Washington, Oregon, and California	0.39
Educational video programs Description of videos on volcanoes with information on obtaining copies.	Worldwide	0.02
Eruption warning and real-time notifications Describes strategy of volcano warnings and the real-time detection of a sudden eruption or lahar and immediate notification of the activity to the public and local, state, and federal emergency-management officials.	Worldwide	0.02
Geologic hazards Links to global information on earthquake, geomagnetic, volcanoes, and landslide hazards plus dynamic maps, images, seismic maps, and geomagnetic data.	Worldwide	0.02
11 other links to worldwide sites on volcanoes	Worldwide	0.02

Is a Bounding Box Good Enough?

The *Browse USGS* application approximates the web page footprints with bounding rectangles. Using bounding polygons would achieve more accurate results, but metadata for all targets would need to contain representations of the bounding x-y strings of their footprint.

Compiling spatial footprints as polygons would require much more work than finding bounding rectangles, particularly for those objects that do not follow well-known boundary sets (States, watersheds, etc.). Moreover, the number of points required to represent a polygon boundary depends on the desired resolution, leaving open the possibility of different metadata for different searches. In contrast, a bounding rectangle is a simple shape with exactly 4 points, regardless of resolution. With few exceptions, a bounding box is clearly recognizable as an approximation, and does not raise questions of precision. While no longer a dominant consideration, due to faster computers, the computational time required to determine the intersection area of polygons is much greater than that of rectangles.

Although using a bounding box will not hurt recall, assuming the box encloses the bounding polygon, the search could be less precise. The bounding rectangle of California, for example, contains all of Nevada. While this is a serious problem when using a Boolean search, a spatial overlay ranking tends to mitigate this effect. When querying for a search within the bounding box of California, a target comprising Nevada would score 0.91; this is high, but still separable from the 1.0 score of targets comprising only California.

The possible consequences of using bounding rectangles instead of polygons can be tested by comparing some common political and natural shapes used in the United States. Table 2 shows that incidental overlaps caused by using bounding boxes result in scores of less than 0.9 in nearly all cases.

Table 2. How elements of some common U.S. political and natural shapes overlay each other.

Geometry	**Maximum spatial overlay score of any 2 elements in the set.**	**Percent of elements which, as query area, have a spatial overlay score >0.9 with at least one other element in their set.**
States	0.91 (CA-NV)	2% (CA)
Counties compared within States	0.97	0.13%
Watersheds (8-digit hydrologic unit codes)	0.96	1.7%

Note: $k_q = 0.1$, $k_t = 0.5$,

Conclusions

A spatial overlay score based upon how well an object's footprint matches the search area, and ranging between 0 (no overlay) and 1 (perfect fit), provides an effective way to spatially search a collection of web pages or reports. This method avoids many of the problems associated with an "in/out" Boolean search. Moreover, spatial overlay ranking appears to work well even when spatial extent is defined only by a bounding box.

References

Dice, L.R. (1945), Measures *of the Amount of Ecologic Association Between Species*, Ecology, Vol. 26, No. 3, July 1945, pp. 297-302.

Hill, L.L.. (1990), Access to Geographic Concepts in Online Bibliographic Files: *Effectiveness of Current Practices and the Potential of a Graphic Interface*, Ph.D. Thesis, Department of Library and Information Science, School of Library and Information Science, University of Pittsburgh, 200p.