Web-Based Delineation of Imprecise Regions

Avi Arampatzis, Marc van Kreveld,
Iris Reinbacher
Institute of Information and Computing Sciences, Utrecht University
{avgerino,marc,iris}@cs.uu.nl

Christopher B. Jones, Subodh Vaid
School of Computer Science, Cardiff University
{c.b.jones,subodh.vaid}@cs.cardiff.ac.uk

Paul Clough, Hideo Joho,
Mark Sanderson
Department of Information Studies, University of Sheffield
{p.d.clough,h.joho,m.sanderson}
@sheffield.ac.uk

Marc Benkert, Alexander Wolff
Department of Computer Science, TU Karlsruhe
{mbenkert,awolff}@ira.uka.de

1. INTRODUCTION

Geographical information tools such as online maps and gazetteers have been increasingly developed and become available on the World Wide Web. Such resources are of interest not only in geographic information systems but also information retrieval systems (i.e. search engines). For example, a set of retrieved documents can be categorized based on their geographical attributes found in texts [6]. However resolving the definition of certain geographical terms is still an active research area. Terms that describe so-called "ill-defined regions" such as the "MidWest" in US or the "Midlands" in the UK are such an example. Such region names may be used in Web searches, and therefore it is useful to know their extent. After delineation, the regions can be stored as geographic features in a database or an ontology.

This paper describes several steps in the derivation of boundaries of imprecise regions using the Web as the information source (see also [4]). We discuss how to obtain locations that are part and locations that are not part of the region to be delineated, and then we propose methods to compute the region algorithmically. Experiments are in progress to analyze how well our approach works.

We identify the following steps:

1. Use the Web to find points (cities, towns) inside the unknown region.
2. Find the coordinates of these points, and a bounding box.
3. Use the bounding box to find coordinates of other cities and points, apparently lying outside.
4. Find a reasonable boundary of the imprecise region using the points inside and outside.

Steps 1 and 4 are the more challenging ones, so we describe these in more detail. Steps 2 and 3 can simply be done using a geographic database or an ontology.

2. EXTRACTING AND LINKING GEOGRAPHIC NAMES USING TRIGGER PHRASES (STEP 1)

Our approach uses a set of text patterns (called trigger phrases) to extract (i) membership of a given area, (ii) short descriptions of geographical names, and (iii) geographical relations between two or more places. Examples of the trigger phrases for the first membership task are as follows (where "MidWest" is a target area and X, Y, Z are the members):

SUCH AS: "MidWest [states|cities|towns] such as X, Y, and Z"
AND OTHER: "X, Y, Z and other MidWest [states|cities|towns]"
INCLUDING: "MidWest [states|cities|towns], including X, Y, and Z"

For example, by simply submitting the trigger phrase "MidWest states such as" as a query to Google, the following member was found in the snippets of the top 10 documents (the numbers are the frequency of occurrence):

Illinois (3)
Indiana (3)
Iowa (2)
Kansas (1)
Michigan (2)
Minnesota (3)
Nebraska (1)
Ohio (3)
Wisconsin (1)
the Dakotas (1)
---
disaster history (1)
environmental data (1)
organization (1)

Although there was some noise, this illustrates a promising approach to provide a set of members in a region. In a similar way, cities and towns in the British Midlands can be identified. It is necessary to study how the boundaries shift when a different cut-off value for frequency of occurrences is used for the selection of members. Some places seem to

gather a greater consensus of being a part of a region than others.

Our approach is simple and efficient, and has been shown to work successfully in, for example, question answering tasks (e.g. [3,5]) where the accurate detection of explicit relations between query and answer strings is required. One of the reasons behind the success of such approaches is due to the use of large amount of texts indexed by a search engine. While the occurrence of trigger phrases can be rare, we only need a couple of matching sentences to extract related names/descriptions.

3. DETERMINING A REASONABLE BOUNDARY OF AN IMPRECISE REGION (STEP 4)

After identifying members in a region, possibly with some noise, we obtain their coordinates using an ontology. Then we compute the bounding box of these points; we consider them to be blue. From these coordinates, we compute a bounding box R, which we enlarge by 20% to get the surroundings of the region of interest as well.

Again using the ontology, we identify locations that lie in the bounding box R, but were not found in step 1. These are likely to be outside the imprecise region. The coordinates of these locations give a set of points that we consider to be red. Now we need to find a region (polygon) that has (nearly) all blue points inside and (nearly) all red points outside. This polygon should have properties such as compact, simply-connected, smooth boundary, etc.

Algorithms to compute such polygons have been proposed before [1], whereVoronoi diagrams are used. In their application the input was assumed to be correct, that is, all colors were correctly assigned. We propose two algorithms for our application, where we cannot assume correct coloring of the points. False positives and false negatives are likely to occur.

The first algorithm starts with an alpha-shape of the blue points [2]. Only the main, big blue component is maintained, the other blue points are outliers (false positives) and are discarded. Then we adapt the polygon to transfer more red points to the outside (if none are inside, we are done). We do this incrementally, while keeping the compact shape of the polygon. We choose the red point closest to the polygon boundary and change the shape. If no red point lies close to the boundary, or the compact shape cannot be maintained, we stop and report the polygon. Red points remaining inside are assumed to be false negatives.

The second algorithm is based on the Delaunay triangulation. We compute the Delaunay triangulation of all red and blue points, and give all edges one of three colors. An edge is blue if both endpoints are blue, an edge is red if both endpoints are red, and an edge is green otherwise. If we connect the midpoints of the green edges around the biggest blue component we get a possible shape for the polygon. To deal with false colors, we define for each point its green angle: it is the largest angle between two incident green edges that has no red or blue edge in between. We incrementally re-color any point whose green angle is larger than some angle A, which must be chosen larger than 180 degrees. Re-coloring a point (red to blue, or blue to red) changes the color of all the incident edges, and often the green angle of its neighbors. We continue this process until all points have green angle at most A. Then we take as the boundary of the imprecise region the connection of the midpoints of the green edges around the largest blue component.

The approaches described in this paper are implemented and experiments will be done to test the methods.

4. ACKNOWLEDGEMENT

This research is supported by the EU-IST Project No. IST-2001-35047 (SPIRIT).

5. REFERENCES

[1] Alani H., C.B. Jones and D.S. Tudhope (2001) "Voronoi-based region approximation for geographical information retrieval with gazetteers". International Journal of Geographical Information Science, 15(4), 287-306.

[2] Edelsbrunner, H., D.G. Kirkpatrick, and R. Seidel (1983). "On the shape of a set of points in the plane". IEEE Transactions on Information Theory, IT-29(4):551-559.

[3] Joho, H. and M. Sanderson, (2000) "Retrieving Descriptive Phrases from Large Amounts of Free Text". In: Proceedings of the 9th International Conference on Information and Knowledge Management, 180-186, McLean, VA: ACM.

[4] Markowetz, A., T. Brinkhoff, and B. Seeger (2003) "Exploiting the Internet as a Geospatial Database". ISPRS WG IV/5 Workshop on Next Generation Geospatial Information.

[5] Ravichandran, D. and E. Hovy, (2002) "Learning Surface Text Patterns for a Question Answering System". In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 41-47, Philadelphia, PA: ACL.

[6] Smith, D. (2002) "Detecting and Browsing Events in Unstructured text". In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 73-80, Tampere, Finland: ACM.