# Geographical Information Retrieval

CHRISTOPHER B. JONES† AND ROSS S. PURVES‡

†School of Computer Science, Cardiff University, 5, The Parade, Cardiff, CF24 3XF, UK. email: c.b.jones@cs.cf.ac.uk

‡Department of Geography, University of Zurich, Switzerland. email: ross.purves@geo.uzh.ch

## 1 Introduction

Geographical information is recorded in a wide variety of media and document types. There are innumerable paper-based books, reports, images and maps; there are computer databases and digital maps along with vast numbers of web pages containing text, images and digital versions of articles, books and reports. Over the past few decades information technology for accessing geographical information has focused on the combination of digital maps and databases that characterise the majority of GIS. It is only in recent years that much attention has been paid to the development of computer systems to retrieve geographically-specific information from the relatively unstructured but immense resource of documents that compose the world-wide web (e.g. Larson 1996, McCurley 2001, Egenhofer 2002, Rauch *et al.* 2003, Purves *et al.* 2007).

Because everything we do takes place in a geographical context it is not surprising that many requirements for information on the web have a geographical focus (it has been estimated that between 13-15% of web queries submitted to traditional search engines contain a place name (Sanderson and Kohler 2003, Jones *et al.* this issue)). It is this need for public access to geographical information on the web that has been a major motivation for the growth of the academic field of geographical information retrieval (GIR).

Much current research in GIR can be regarded as an extension of the field of information retrieval (IR) (Baeza-Yates and Ribeiro-Neto 1999), and indeed has primarily been undertaken by researchers from the IR rather than the GIScience research community. The products of work in IR have provided the foundation for current web search engine technology, with its focus on access to web documents that are relevant to user queries taking the form of a phrase or set of key words. This latter type of query is in contrast to the more formal approach common in GIS, where specific geo-referenced data objects are retrieved from a structured database in response to queries that can stipulate precise spatial constraints. IR methods usually identify web pages that contain query terms and rank documents using statistical methods that are intended to highlight the most relevant. For some forms of geographical search, particularly when looking for common resources within relatively large geographic extents (e.g. hotels in London), this approach can work but it is fraught with limitations.

From the user's perspective these limitations can be manifested in a failure to distinguish between different instances of the same place name, a lack of ability to deal with spatial qualifiers such as "near" or "north", a lack of methods to rank and explore results with respect to their geographic relevance and the non-retrieval of resources which are geographically relevant but use a place name different from that specified in the query. GIR is therefore concerned with improving the quality of geographically-specific information retrieval with a

focus on access to unstructured documents such as those found on the web. There are several aspects of GIR that need to be improved and that constitute significant research challenges. These include the problems of

- detecting geographical references in the form of place names and associated spatial natural language qualifiers within text documents and in users queries;
- disambiguating place names to determine which particular instance of a name is intended;
- geometric interpretation of the meaning of vague place names, such as the "Midlands" and of vague spatial language such as "near";
- indexing documents with respect to their geographic context as well as their non-spatial thematic content;
- ranking the relevance of documents with respect to geography as well as theme;
- developing effective user interfaces that help users to find what they want;
- and developing methods to evaluate the success of GIR.

In the following we elaborate briefly upon these issues before introducing the content of the articles within this special issue.

## 2 Issues in GIR

### *2.1 Detecting geographic references*

Place names (or toponyms) can be used to refer to places on Earth but they also occur frequently within the names of organisations and as part of people's names. It is also the case that place names may be used metonymically, for example to refer to administrative entities as in "talks with Washington". The process of geo-parsing is concerned with analysing text to identify the presence of place names and other spatial language and distinguishing the genuine geographical occurrences of place name usage from those where they are being used to refer to some other entity. This process is often treated as an extension of Named Entity Recognition that is a standard part of linguistic analysis in Natural Language Processing (NLP). The task of detecting genuine geographic references is addressed in this issue by Levelling and Hartrumpf and by Stokes *et al*.

### *2.2 Disambiguating place names*

Once it has been established that a place name is being used in a geographic sense, the problem remains of determining uniquely the place to which the name refers. There are many names that are shared between different places, for example Richmond, Newport and Springfield. When a human reads a document with a place name in it they will tend to resolve ambiguity using knowledge gained from contextual clues within the document. Automatic resolution of geographic scope, and hence disambiguation of place names, attempts to mimic the methods humans use, for example by considering together all of the place names in the document. If a place name occurs in association with a set of other names, several of which are neighbouring places or are instances of places within the same parent region, then that provides evidence to distinguish which meaning is implied. Equally if the text mentions a parent or child region of an instance of the name then that can help to determine the particular sense that is intended.

Several of the articles in this special issue address the issues of disambiguation (Buscaldi *et al*, Overell and Ruger, Stokes *et al*).

### *2.3 Vague geographic terminology*

Many of the place names that users employ when searching on the web are of an informal or vernacular nature, often without precise boundaries. Examples include the Pennines and the Borders in the UK, the South of France, the Swiss Mittelland and the Midwest in the USA. Existing geographical search facilities make use of place name resources, typically gazetteers, that are based largely on the administrative names of places and which do not, in general include any representations of vernacular names (Hill 2006). There have been a few studies of methods of determining what may be a fuzzy extent for such places (e.g. Montello *et al.* 2003) but there is still much to be done. An approach adopted by Arampatzis *et al.* (2006) and Jones *et al.* (in press) is to use web harvesting methods that can determine associations between a vernacular place and the names of other places within it or in the vicinity. Spatial extent can then be modelled using for example density surfaces or Delaunay triangulation based methods.

The spatial language, such as near, close, between and north of, that accompanies place name terminology can be as vague as some of the place names. Being able to interpret such terms will help in analysing the geographic context of documents and in interpreting user queries that employ vague spatial language. There have been previous studies of the meaning of natural language qualitative spatial relations (e.g. Worboys 2001) and in this issue Schockaert *et al.* analyse how nearness is described using phrases such as “walking distance” based on descriptions found on the web of places in the vicinity of hotels.

### *2.4 Spatial and textual indexing*

When web documents have been categorised according to their geographical context they must be indexed in a way that enables them to be found quickly in response to user queries. Techniques for indexing documents according to the words that they contain are well established. Typically an inverted file of documents is created in which each word is associated with a list of the documents that contain the word. This text indexing can be combined with a spatial index that records which documents relate to particular regions of space. Building a spatial index of documents can be done if each document has one or more document footprints that represent the regions of geographic space to which the document refers. Each document footprint may correspond to the spatial extent of a geographic reference that occurs in the document. If there are many such references an effort may be made to establish the main geographic foci of the document as represented by a smaller number of footprints (Amitay *et al.* 2004, Wang *et al.*, 2005, Silva *et al.* 2006). These footprints can then be indexed in the same way that any other piece of geometry would be indexed in a conventional GIS. A challenge remains however to find efficient ways to combine text and spatial indexes (see Vaid *et al.* 2005 for a summary of some approaches to spatio-texual indexing).

Retrieval of relevant documents requires matching the query specification to the characteristics of the indexed documents. As mentioned above, in conventional web search engines this starts by finding those documents that contain the query terms, before ranking the resulting documents. For geographical search there is a need to match the geographical component of the query with the geographical context of the documents as represented by the document footprints. GIR queries can be characterised as a triplet of <theme><spatial relationship><location> composed of a topic of interest in combination with a place name qualified by a spatial preposition such as near, in, or north of. The combination of place name (after disambiguation) and spatial preposition can be used to generate a query footprint representing, for example, the interpretation of an expression such as “near Bristol”. This query footprint can then be used to access the relevant part of the spatial index and hence find document footprints that intersect the query footprint. The retrieved documents will then be the members of this latter set of geographically relevant documents that also contain the thematic

(text) query terms in the query. The contribution of Frontiera *et al.* to this special issue investigates the use of probabilistic methods for assessing the spatial similarity of document and query footprints.

### *2.5 Geographical relevance ranking*

Having found a set of potentially relevant documents they should then be ranked by some measures of their estimated degree of relevance to the query. Relevance with respect to the thematic part of the query and the retrieved documents can, for example, be represented by a score that takes into account factors such as the frequency of occurrence of query terms within retrieved documents as a function of the overall frequency of query terms within the whole collection. The spatial score can be some measure of the geometric match between the query footprint and the document footprints. These two scores can then be combined to find an overall relevance. Kreveld *et al.* (2005) have described some methods for doing this in which the text and spatial scores are normalised to values between 0 and 1 before calculating their distance from an ideal combined score in the two-dimensional space of text and spatial scores.

### *2.6 User interfaces*

Retrieval of documents using GIR also provides a variety of research challenges in the development of user interfaces. Query formulation generally requires, as indicated above, specification of a triplet of <theme><spatial relationship><location>. This can easily be facilitated by a simple structured interface, but this assumes sufficient geographic knowledge to specify relevant place names. Other approaches to query formulation allow users to sketch a region of interest on a map, and enter a related concept within a textbox (Purves *et al.* 2007). This approach presupposes that users are interested in a spatial relationship representing containment, but to date very little progress has been made in developing map-based interfaces that allow users the option of graphically specifying other spatial relationships.

The results of a query to a GIR system can be treated in an identical manner to those of a traditional search engine, and simply displayed as a ranked list. In practice, the nature of geographic search and the pervasiveness of map-based web services mean that the overlaying of results on a map has become a natural and expected visualisation mechanism. Several challenges exist in displaying the results of GIR searches through a map interface. Many relevant documents may have the same geographic footprint, and techniques are required to allow the aggregation of these relevant documents, whilst summarising or filtering duplicate content. Equally, document footprints often have scopes which are not sensibly represented at all scales as a point (e.g. London) and methods allowing users to explore documents with extensive geographic scopes in meaningful ways are required. Finally, there is much room for techniques from the Geovisualisation community to be applied in exploring the typically large sets of documents that are returned by GIR.

### *2.7 User studies and evaluation*

In developing GIR techniques it is very important to take cognisance of user needs and to develop techniques to evaluate the quality of approaches to GIR. With respect to user needs, work is required to analyse where gaps exist in current approaches and to analyse query logs to assess where users fail or have difficulty in search. Unfortunately at present there is a lack of availability of real query logs for academic research. The release of a query log by Microsoft as part of the GeoCLEF campaign of 2007 illustrates the importance of this issue in commercial search as does the study by Jones *et al.* in this issue.

Evaluation is a key part of IR. In general, evaluation strategies within IR either take the form of system-focused or user-centered studies (Spark Jones and Willet 1997). The former are based around the use of test collections, the measuring of the relevance of documents to specific queries and the calculation of standard IR measures such as precision and recall for a variety of systems and settings. Such methods require substantial resources to implement, particularly since relevance judgments must be performed manually and the numbers of judgments required are large. Within IR, the long-running TREC experiments have provided a mechanism for the pooling of resources to the mutual benefit of the research community (Voorhees and Harman 2000). In GIR, GeoCLEF (Gey *et al.* 2005, Gey *et al.* 2006) has gone some way towards addressing this need by mounting a campaign to compare performance of a range of approaches to GIR across a set of multilingual queries performed on document collections based on newspaper articles. However, there is much room for further research on relevance judgment and measures for GIR. Current techniques are based on a single dimension of relevance using standard IR measures, which in general do not take account of the G (geography) in GIR (Purves and Clough 2006). To date, very little user-centered evaluation has been performed in GIR.

## 3 The articles in this issue

In an effort to understand more about the use of geographical queries on the web, Jones *et al.* analyzed a log of queries to a commercial web search engine. They identified the presence of place names and associated spatial qualifiers in conjunction with the likely locations of the users as indicated by their IP (Internet Protocol) addresses. The study found that about 13% of queries contained place names, a figure that is similar to that in an earlier study by Sanderson and Kohler (2003). They looked at the distances between users' location and the places for which they searched for information and found that people are, perhaps unsurprisingly, most likely to search for information local to themselves and information about cities is sought more frequently than that about countries. When considering search topic they found correlations between different types of phenomena and the distances to the places of interest. Restaurant searches for example tended to be nearer to the user location than those for hotels. Such knowledge could be used to estimate the meaning of "near" when used in conjunction with different types of phenomena.

The authors also looked at the way in which users reformulate queries distinguishing between changes due to corrections in spelling and changes between the location itself, which provides important information on the variability of geographic needs. The study revealed differences in the degree of change according to user location. Users in California were more likely to confine changes in their search to shorter distances from their location than those in Vermont, presumably reflecting the difference in density and spatial arrangement of facilities between these two states. In this case it is possible to envisage adapting responses to user queries according to their place of origin.

Stokes *et al.* take a critical look at natural language processing (NLP) based approaches to the detection and disambiguation ("resolution" in their terminology) of toponyms and hence their annotation within documents. Using a manually annotated GeoCLEF dataset to evaluate some standard NLP methods, they found that their performance was lower than previously publicised. They also found, for their experiment, that a simple detection approach employing a gazetteer to recognise the presence of toponyms outperformed the NLP methods. Their study looks at the interaction between errors in the two stages of processing, finding that in some situations errors in detection can be negated at the disambiguation stage if the latter cannot recognise the identified name with its toponym knowledge resources.

When looking at the effects of these NLP pre-processing methods on the results of geographical information retrieval experiments, a notable conclusion was that their best performing NLP method was only marginally superior to a method that did not employ geographical annotation methods at all. The authors also address the issue of geographical query expansion whereby a query is supplemented with additional toponyms that can be regarded as equivalent in some sense to that in the query. They introduce and demonstrate the benefits of a similarity normalisation method that overcomes problems that can arise when the expansion results in very large numbers of additional query terms. They discuss the effects of query expansion and highlight the potential to improve GIR performance using more sophisticated approaches, taking account of both geo and non-geo terms.

Overell and Ruger also address the problem of disambiguation of place names, by exploiting prior knowledge of the associations, or co-occurrences, between place names in existing documents. They use the text and links of Wikipedia articles to build a co-occurrence model of place names which records successive occurrences of links to place names within individual articles. To construct the co-occurrences model the place names in Wikipedia have themselves to be disambiguated, but because of the quality of metadata associated with the articles this is a simpler problem than that of disambiguation within text documents in general. The co-occurrence model is richer than a conventional gazetteer in that it records knowledge of the order in which place names occur within multiple documents. This knowledge allows, for example, the construction of rules of association between place names, based on the substantial resource of Wikipedia. Thus if a particular instance of a name such as “London” is followed by the name of its parent province, “Ontario” then that can be used to infer that it is the Canadian London rather than the British one. The authors apply their methods to GeoCLEF experiments, and find that there is no significant improvement in performance as a result of their method of disambiguation. However, they also comment that this may be in part due to the nature of the queries tested and suggest that improvements will occur for some types of queries as methods of integrating the spatial and textual indexes are improved.

The paper by Levelling and Hartrumpf advances the subject of toponym detection by focussing on the problem of metonymic use of place names in which place names are used as a figure of speech and refer not to the place itself but rather some related meaning in which the location is a shorthand for, for example, an event, government, organisation or product. The study is based on German language documents and uses a machine learning approach to recognise metonymic occurrences, based on training data some of which has been obtained with a disambiguating parser that employs linguistic methods to detect situations such as that in which a location name occurs where a person or an institution should appear as the subject of a verb. These linguistic methods are appropriate for use in obtaining training data but can be too slow for general purpose analysis of very large quantities of data. The machine learning classifier results in accuracy of about 80% in identifying literal uses of place names. When subsequently applied to indexing of place names for experiments retrieving geographical information from GeoCLEF, the results did provide some improvement in precision.

Buscaldi *et al.* also address the problem of toponym disambiguation, by measuring similarity between a toponym to disambiguate and the toponyms within the local context of the occurrence of the word. They applied the measure of Conceptual Density (CD) which exploits knowledge of word relationships encoded within the lexical database Wordnet (Fellbaum, 1998). In the authors’ version of CD they construct a local subhierarchy for the word to be disambiguated based on part-of relationships. They also make use of the Wordnet synsets of the word (i.e. a set of words with similar meaning) to disambiguate its toponym neighbours within the textual context. CD is then a function of the numbers of synsets of the toponym and its

context toponyms that are found within the subhierarchy. The approach was compared to other methods including that of simply assigning the most frequent sense of a toponym (so for example “London” would be assumed to be the London that is the capital of the UK). Their CD approach showed some benefits using particular measures of performance but overall the most frequent sense methods proved to be superior. The result however was considered by the authors as inconclusive since they argue that Wordnet is a relatively poor source of place name knowledge.

Schockaert *et al.* address the issues of vagueness that arise in the spatial language that people use when making queries (near to, within walking distance, etc) and the vagueness in the spatial extent of some geographical regions such as neighbourhoods within cities. They employ fuzzy methods to model interpretations of spatial language prepositions and of the extent of vague places. In both cases the parameters for these models are based on data obtained from web pages that include instances of the various types of vague language and vague neighbourhoods. Thus web sites that describe hotels often include textual descriptions that describe their location using phrases such as “within walking distance” of some prominent location such as a well known local landmark. The spatial language expressions (or hints as they are referred to) can then be compared to ground truth knowledge of the actual distances involved. The set of actual distances that correspond with the use of a particular expression can be used to create a fuzzy set membership function.

In the case of vague neighbourhoods, web documents are used to find places that are likely to lie inside the neighbourhood. This is done by issuing a query to a local search engine naming the vague place as a query term along with a definite parent region (which could be a city name). Georeferenced named features such as restaurants that include the vague name, or which are associated with descriptions that mention the name, are retrieved as candidate contained places. An assumption is made about the confidence with which these candidate places actually lie inside the target vague place, and a fuzzy membership function is constructed with a definite core region and a zone of possible containment.

The contribution by Frontiera *et al.* is in the area of geographic relevance ranking. They focus specifically on the spatial aspects of relevance ranking looking at how documents can be ranked as a function of the degree of spatial similarity between the document footprint and the query footprint. Existing methods of spatial relevance ranking are usually based on a single measure of similarity such as, for example, the degree of overlap between the query and document footprints divided by the area of the query footprint. The paper’s novelty lies in introducing probabilistic measures of spatial relevance to the problem through the method of logistic regression. This is able to take into account multiple measures of similarity and to do it in a way that exploits a training dataset to determine the values of logistic regression coefficients for each of the similarity measures. The similarity measures used are overlap between query and document footprint divided by the query footprint for one measure, and by the document footprint for the second. A measure is also introduced that is related to the proportions of the query and document footprint respectively that are on land. The test collection is specific to the research work, consisting of footprints of various regions in California, some of which correspond to well defined administrative areas while others are less precisely defined with only bounding boxes. They show that with their method, using only bounding boxes, as opposed to polygons or convex hulls, they can outperform a variety of existing published methods in terms of both precision and recall.

Overall, we believe that the set of papers presented in this special issue provides a snapshot of current research within the field of Geographical Information Retrieval. We hope that they will not only prove

interesting to the readers of IJGIS, but stimulate contributions from the GIScience research community in this important and rapidly growing research field.

**References**


AMITAY, E., HAR'EL, N., SIVAN, R. AND SOFFER, A., 2004, Web-a-where: geotagging web content. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR04)*, 25-29 July 2004, Sheffield, UK (New York: ACM Press), pp. 273–280.

ARAMPATZIS, A., VAN KREVELD, M., REINBACHER, I., JONES, C. B., VAID, S., CLOUGH, P., JOHO, H. AND SANDERSON, M., 2006, Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, **30**, 436–459.

BAEZA-YATES, R AND RIBEIRO-NETO, B., 1999, *Modern Information Retrieval* (Boston, MA: Addison Wesley).

EGENHOFER, M. J., 2002, Toward the semantic geospatial web. In *Proceedings of the 10th ACM International Symposium on Geographic Information Systems*, 8-9 November 2002, McLean, VA, USA (New York: ACM Press), pp. 1–4.

FELLBAUM, C., 1998, *Wordnet: an electronic lexical database* (Cambridge, MA: MIT Press).

GEY, F., LARSON, R., SANDERSON, M., JOHO, H. AND CLOUGH, P. (2005) GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track. *Working Notes for the Workshop on the Cross Language Evaluation Forum 2005* - CD-ROM.

GEY, F., LARSON, R., SANDERSON, M., BISCHOFF, K., MANDL, T., WOMSER-HACKER, C. SANTOS, D., ROCHA, P., DI NUNZIO, G. M., AND FERRO N., 2006, GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. *Working Notes for the Workshop on the Cross Language Evaluation Forum 2006* - CD-ROM.

HILL, L. L., 2006, *Georeferencing: The Geographic Associations of Information* (Cambridge, MA: MIT Press).

JONES, C.B., PURVES, R.S., CLOUGH, P.D AND JOHO, H., (In press), Modelling Vague Places with Knowledge from the Web, *International Journal of Geographical Information Science*.

KREVELD, M. VAN., REINBACHER, I., ARAMPATZIS, A. AND ZWOL, R. VAN., 2005, Multi-dimensional scattered ranking methods for geographic information retrieval. *Geoinformatica*, **9**, 61–84.

LARSON, R., 1996, Geographic information retrieval and spatial browsing. In *GIS and Libraries: Patrons, Maps and Spatial Information*, L. Smith and M. Gluck (Eds.) (Urbana-Champaign: University of Illinois), pp. 81–124.

MCCURLEY, S. K., 2001, Geospatial mapping and navigation of the web. In *Proceedings of the 10th International WWW Conference*, 1-5 May 2001, Hong Kong (New York: ACM Press), pp. 221–229.

MONTELLO, D., GOODCHILD, M., GOTTSEGEN, J. AND FOHL, P., 2003, Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition and Computation*, **3**,185–204.

PURVES, R. S. AND CLOUGH, P., 2006, Judging spatial relevance and document location for geographic information retrieval, extended abstract. In *Proceedings of 4th International Conference on Geographic Information Science (GIScience 2006)*, 20-23 September 2006, Muenster, Germany (Muenster : IfGI Prints), pp. 159–164.

PURVES, R. S., CLOUGH, P., JONES, C. B.,ARAMPATZIS, A., BUCHER, B., FINCH, D., FU, G., JOHO, H., SYED, A. K., VAID, S. AND YANG, B., 2007, The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, **21**, 717–745.

RAUCH, E., BUKATIN, M. AND BAKER, K., 2003, A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, May 31 2003, Edmonton, Alberta, Canada (Morristown, NJ, USA: ACL Press), pp. 50–54.

SANDERSON, M AND KOHLER, J., 2004, Analyzing geographic queries. In *Proceedings of the 2004 Workshop on Geographic Information Retrieval*, 29 July 2004, Sheffield, UK. Available online at: http://www.geo.uzh.ch/~rsp/gir/abstracts/sanderson.pdf (accessed 6 August 2007).

SILVA, M., MARTINS, B., CHAVES, M., CARDOSO, N. AND AFONSO, A. P., 2006, Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, **30**, 378–399.

SPARCK JONES, K AND WILLETT, P. (Eds.), 1997, *Readings in Information Retrieval* (San Francisco: Morgan Kaufmann).

VAID, S., JONES, C. B., JOHO, H. AND SANDERSON, M., 2005, Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th International Symposium on Spatial and Temporal Databases*, 22-24 August 2005, Angra dos Reis, Brazil (Berlin: Springer), pp. 218–235.

VOORHEES, E. M. AND HARMAN, D., 2000, Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management* **36**, 3–35.

WANG, C., XIE, X., WANG, L., LU, Y. AND MA, W., 2005, Detecting geographic locations from web resources. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, 4 November 2005, Bremen, Germany (New York: ACM Press), pp. 17–24.

WORBOYS, M. F., 2001, Nearness relations in environmental space. *International Journal of Geographical Information Science*, **15**, 633–651.