Geographic Information Retrieval: Are We Making Progress?

ROSS PURVES
Department of Geography
University of Zurich
Email: rspurves@gmail.com

What is spatial search? And perhaps more importantly, what are the implications of developing systems which allow some form of spatial search? In this position paper I aim to explore these questions from the perspective of Geographic Information Retrieval, a term originally coined by Larson (1996) in his seminal paper in 1996 and, refined to more specifically refer to unstructured information, primarily as found in text documents, in an editorial written by Chris Jones and myself in a special issue of IJGIS published in 2008 (Jones and Purves, 2008). In that editorial we set out what we believed to be key challenges for GIR. These challenges were based on our experiences in developing the SPIRIT system, one of a number of early GIR systems (e.g., Chen et al., 2006; Lieberman et al., 2007) which sought to bring some form of geographic intelligence to information retrieval. As such, the challenges closely mimicked what believed to be the necessary components of such a system, and put little emphasis on, rightly or wrongly, developments in areas such as linked data and the semantic web.

In the following, I will briefly list these challenges, and attempt to explore progress by looking at the literature from two viewpoints. Firstly, I am interested in the broad range of methodological and disciplinary approaches applied to the different challenges. Does lack of attention to a challenge imply that we simply identified the wrong questions, or that progress in other areas is required before such challenges can be met? Secondly, since spatial search and GIR have now become mainstream in computer science, I am interested in what useful ideas, if any, GIScience and geography might have to offer in this research field.

Challenges and Progress

Detecting geographical references in the form of place names and associated spatial natural language qualifiers within text documents and in users’ queries

When we wrote our editorial, the focus of most work in GIR was on text documents published on the web, of variable lengths, and often containing rich geographic information. Extracting locations from documents had been rightly identified as an important challenge, and the focus here was not per se on what is known as toponym resolution (identifying a single, unique referent for a given place name), but rather toponym recognition (identifying whether a token or phrase within a document can be treated as a spatial referent) (e.g., Leidner, 2008). This area of research had already been the subject of attention from computer linguistics as part of the more general task of Named Entity Recognition. Queries can be seen as forerunners to the processing of other very short documents, such as Tweets, as the advent of georeferenced feeds has provided a perfect opportunity to experiment with, for example, approaches based on machine learning (e.g., Roller et al., 2012). Indeed, many approaches based on machine learning conflate toponym recognition and resolution
and remove the need for both gazetteers and heuristics based on the ways in which place names are used. GIScience has seen increasing attempts to use corpora such as Twitter as sources of information about spatial language—however the extent to which, for instance, some of the ideas of naïve geography (Egenhofer and Mark, 1995) and the related use of spatial language has informed this research appears limited.

Disambiguating place names to determine which particular instance of a name is intended

Having identified candidate referents in documents, a key task is their disambiguation, and thus toponym resolution. Humans typically excel at this, by using both clues within the document and other contextual information, to the extent that booking a flight in error to Sydney, Nova Scotia rather than Australia is the subject of a news story (http://news.bbc.co.uk/2/hi/uk/2172858.stm). However, this task remains a challenging one for automated systems, which typically adopt relatively simple approaches, such as using ancillary information mined from other content (e.g., population counts or co-occurrence with other terms) or simple geometric measures to calculate distances to the nearest unambiguous referent identified (Smith and Crane, 2001; Buscaldi, 2011). However, many of these approaches assume that geography is random (i.e., that ambiguous toponyms have no spatial autocorrelation) despite clear evidence that this is not always the case (Brunner and Purves, 2008). Furthermore, most approaches focus on toponyms related to settlements (Leidner, 2008) despite the importance of dealing with more fine-grained toponyms referring to, for example, individual locations along the route of a hike.

Geometric interpretation of the meaning of vague place names, (e.g., “Midlands”) and of vague spatial language (e.g., “near”)

Many toponyms are not found in official gazetteers. Often, these toponyms are also vague, and in fact seemingly well-defined toponyms may be being used in natural language in a vague sense. In addition, the use of spatial language can take seemingly well-defined concepts and render them vague. Such concepts have moved from being interesting ideas for GIScientists to consider (e.g., Montello et al. 2003; Jones et al. 2008a) to being of fundamental importance in resolving queries successfully. If around 13% of queries really are in some way spatial (Jones et al., 2008b), it is likely that a sizable portion of these will refer to either vague (and vernacular) place names as well as vague spatial language. Thus, considerable effort has been expended on identifying the geometric regions associated with such place names (e.g., Grothe and Schaab, 2009; Keßler et al., 2009) though often the notion of vagueness has been discarded along the way (e.g., Jones et al., 2008a). Indeed, typical search engine representations of near seem to simply assume that metric distance is a “good enough” method for ranking, and base this on point-based models of geographic entities. Finding place names which are not already in gazetteers, and associating these with either metric or topological models of space is an area of research with much potential, especially if it can be scaled up to be relevant for large digital corpora.

Indexing documents with respect to geographic context and non-spatial thematic content

The notion of indexing is central to all efficient search. Early research in GIR explored the influence of different approaches to index construction (e.g., text vs. space-primary) (Vaid et al., 2005) though
in practice often simple approaches based on multiple indexes and intersection operations appear to have been effective. However, three other important questions can be posed with respect to indexing for spatial search. (1) What should be indexed—should a document be represented as a “bag of points” corresponding to all resolved locations found in the text, as a single footprint, or through some other intermediate representation. (2) The approach taken to spatial indexing (e.g., space or object-directed indexing) has important implications for the questions that can then be explored through the corpus, beyond basic retrieval operations. Such questions are also where potential benefits for geography lie; for example if we wish to map corpora onto existing spatial data, then space-directed indexes ensure that a complete tessellation of space, such that all locations can be linked to content at some given (not necessarily equivalent) granularity. (3) It may be that queries and documents could usefully be represented in different ways. Even where document footprints are stored as simple points, it makes sense to use the observation “topology defines, metric refines” and allow query footprints to take more complex forms, enabling for example containment queries to be meaningfully formulated.

**Ranking the relevance of documents with respect to geography as well as theme**

Relevance of documents with respect to spatial search has two obvious components – thematic and geographic relevance. Producing a meaningfully ranked list of documents therefore requires that both of these components are considered. Many commercial search engines appear to treat these as independent variables, displaying documents which are thematically relevant on a map, with geographic relevance either being treated as binary variable, or sometimes ranked according to distance. Kreveld et al. (2005) explored how geographic and thematic relevance could be considered with respect to some ideal score, effectively representing the two relevance dimensions as orthogonal. A further important aspect of ranking is diversity—in the case of spatial search it makes sense to consider how cluster documents are, and attempt to present the user with not only relevant, but geographically (and thematically?) relevant results (Tang et al., 2010). It is also possible to consider incorporating more context into producing ranked sets of results, in particular with respect to mobile search, where user behaviour may allow, for example, predictions to made about future movement (Mountain and Mcfarlane, 2007) or relevance ranking methods which more fully take into account geographic context in a holistic way (Cai et al., 2007; De Sabbata and Reichenbacher, 2012). Interestingly, there seems to be little formal cross-over between work on spatial cognition and relevance ranking in GIR, despite the obvious potential of considering, for example, notions of salience in developing relevance ranking methods.

**Developing effective user interfaces that help users to find what they want**

User interfaces for geographic search have developed little beyond the original prototypes of SPIRIT in the early 2000s. Query formulation generally supposes a triplet of the form <theme><spatial relationship><location> which can be achieved either by use of a simple structured interface, a multi-modal interface (using a query box to specify theme, the map extent to imply location and typically ignoring spatial relationships), or by processing free text entries to extract some or all of the components of the above triplet. Equally, and perhaps even more surprisingly, results display in
spatial search has not progresses much beyond the display of points on maps, despite some attempts to develop more novel query display approaches (e.g., Hobona et al., 2006; Carmo et al. 2007). Again, there is little evidence of meaningful crossover from the geovizualization community directly dealing with search, as opposed to exploration, tasks.

**Developing methods to evaluate the success of GIR**

When we wrote our editorial the importance of developing effective approaches to evaluation was very clear for many researchers, and indeed the GeoCLEF campaign (e.g., Mandl et al., 2008) focussed on a system-centred strategy to allow comparison between systems. However, this campaign was largely unsuccessful in demonstrating the benefits of GIR over and above simple tweaks to textual systems. However, evaluations of individual GIR systems have often shown benefits over simple textual baselines. The obvious question that one can then pose is why? I believe a number of aspects particular to GIR have been neglected here. Firstly, corpora typically have very specific properties, with for example target audiences having a strong influence on the spatial language, and in particular the toponyms, used. Secondly, evaluating geographic relevance, especially at a local scale appears to be much more challenging than thematic (e.g., Clough et al., 2006, Ostermann et al., 2013), often requiring detailed local knowledge. Traditional and emerging approaches to evaluation based on pooled judgements and crowd sourcing are ill-suited to judging such nuances, and the need for more qualitative user-centred evaluation for judging long-tail queries should not be underestimated.

A second key point in developing effective evaluation strategies concerns harnessing the information which can be obtained from query logs in order to better define how users actually interact with systems. Work by Jones et al. (2008b), importantly performed under the auspices of a search engine company, suggested considerable potential for such explorations. The controversial release of the AOL query logs also demonstrated the potential for analysing the properties of geographic queries in a traditional search engine (Henrich and Lüdecke, 2007), however outside the commercial world access to such data remains rare. Better understanding how users formulate geographic queries, and linking this to basic notions of spatial cognition seems to me to be an obvious area where interdisciplinary research could rapidly bring tangible benefits.

**Developments**

Perhaps the most obvious development which could influence spatial search is the parallel rise of mobile devices and the use of social media. Mobile devices have important implications for context and the way it is used in geographic search, but most applications have continued to focus on the use of well structured (e.g., point of interest) data, and the integration of unstructured data beyond mobile versions of traditional search mechanisms has seen surprisingly little development. The rise of social media has resulted in a need for methods to index and query what are effectively very short documents, and approaches based on machine learning dominate here. A further, perhaps less obvious development is the increase in the open availability of large corpora of legacy documents (e.g., [http://chroniclingamerica.loc.gov/](http://chroniclingamerica.loc.gov/) or [http://textberg.ch/site/en/welcome/](http://textberg.ch/site/en/welcome/)) often with a specific geographic or thematic focus. The potential of such corpora for answering geographic questions has
been, in my view, neglected, and I believe some of the greatest potential for interdisciplinary collaborations between researchers.

In looking back at the challenges we set, a few points emerge. Toponym recognition and resolution remain absolutely fundamental, and despite Leidner’s pleas (Leidner, 2006), gold standards allowing us to evaluate and compare methods have not emerged. Indexes, though interesting, have seen much less research than some areas, at least in part because simple approaches appear to often be “good enough.” By contrast, dealing with vague and vernacular placenames was, and remains, a hot topic, despite the fact that clear evidence of concrete benefits for retrieval task (as opposed to assertions) perhaps remain illusory. Relevance ranking has remained an area which has received relatively little attention, despite the seemingly clear importance of presenting results which are spatially divergent, and indeed the presentation of such results has obvious links with the interfaces of systems providing GIR. Finally, evaluation, despite being the subject of significant attention through the GeoCLEF series remains an area with much potential, where we can perhaps more effectively learn and pool research results.

Acknowledgements: This is a position paper, and the views asserted here are mine and mine alone. Nonetheless, these views are the result of a long and valuable collaboration with many colleagues, and I would like to acknowledge in particular my discussions with Paul Clough, Curdin Derungs, Alistair Edwardes, Chris Jones and Mark Sanderson.

References


