Workshop on Geographic Information Retrieval, SIGIR 2004

Ross Purves
Department of Geography
University of Zürich
Switzerland

rsp@geo.unizh.ch

Chris Jones
School of Computer Science
Cardiff University
Wales, UK

c.b.jones@cs.cardiff.ac.uk

Geographic Information Retrieval is fast emerging as an interdisciplinary hot-topic, both in an academic and commercial sense. Retrieving data based not only on conceptual key words, but some notion of the locational relevance of the information requires research of a range of techniques, for example

- the extraction of geographic terms from structured and, more challengingly, unstructured data;
- the identification and removal of ambiguities in such extraction procedures;
- methodologies for efficiently storing information about locations and their relationships;
- development of search engines and algorithms to take advantage of such geographic information;
- the combination of geographic and contextual relevance to give a meaningful combined relevance to documents; and
- techniques to allow the user to interact with and explore the results of queries to a geographically-aware IR system.

The idea for this workshop germinated, as a result of the organisers' involvement in a project focussing on Geographic Information Retrieval (Spatially Aware Information Retrieval on the Internet or SPIRIT). Our project involves researchers from information retrieval, computation geometry, geographic information science, national mapping agencies, computing science and cartography. Such a diverse group reflects the interdisciplinary nature of research in this field, and we decided that organising a workshop within the auspices of SIGIR 2004 would be an ideal opportunity to bring together the growing community of researchers and practitioners working in the field of geographic information retrieval (GIR) to both discuss progress within the field and identify future research strands.

The workshop was divided into four sessions, entitled *GIR Systems*, *Problems in GIR*, *Extracting Geographic Information from Unstructured Text* and *Test Collections* respectively. The intention was to provide an opportunity to describe some applications of Geographic Information Retrieval before working through a set of specific research themes and finally closing the workshop with a discussion of the possibilities for joint outcomes from the workshop.

The session on ***GIR systems*** opened with a talk from Sanderson and Kohler entitled *Analysing geographic queries.* They presented the results of analysis of a log of queries submitted to the Excite search engine with a view to understanding the frequency and nature of geographically-specific queries. They found that about one fifth of all queries submitted were geographical as determined by the presence of a

geographical term such as a place name, a post code, a type of place or a directional qualifier such as north.

The second paper in the session - *Yellow page driven methods of collecting and scoring spatial web documents* - described a complete GIR system implemented by Sagara and Kitsuregawa. This paper explained aspects of the system for retrieving and scoring geographically specific documents from the web with a prototype spatial search engine. They used yellow pages to generate key words to find documents on the web relating to listed businesses. These were then scored, according to measures of popularity and reliability, and spatially indexed within the web search engine.

Gey and Carl, in a paper entitled *Geotemporal access to multilingual documents* addressed the issue of indexing Russian news articles with respect to space and time, for use in a map-based query and retrieval system. A major issue was the transliteration of Romanized gazetteer versions of place names with the Cyrillic names used in the original texts – merging such schemes across spatial datasets is a major challenge in multilingual GIR.

Given a set of coordinates how can they be transformed to appropriate textual names? Naaman et al. used digital map data to determine the containing regions of the coordinates of digital photographs and hence assign names based on importance criteria in a paper called *Assigning textual names to sets of geographic coordinates*. They also linked the resulting place name directionally to major nearby cities to provide context (e.g. 40kms NE of San Francisco). An important part of this work was an initial evaluation of the system's success in assigning place names, which showed that the algorithms implemented were generally successful.

The final paper in this session, *Information mining: extracting, exploring and visualising geo-referenced information*, from Chrisment et al. described the use of techniques based on data mining and exploration to aid users in analysing and visualising geo-referenced data. They gave an example of a case study where data based around the country of the authors' affiliation was used to explore publication data from their institution. Such techniques, which exploit large multi-dimensional datasets are likely to increase in importance in GIR.

The second session was titled ***Problems in GIR*** and opened with a paper from Arampatzis et al. entitled *Web-based delineation of imprecise regions*. This paper first used the web to identify places which were related with a region whose borders were spatially ill-defined, for example the south of France. Locations associated with such an imprecise region were geo-referenced, and a point dataset of candidate locations constructed. This set of candidate locations was then used to delineate a polygon based boundary for the imprecise region.

The second paper in this session, *Ranking and representation for GIR*, from Larson and Frontiera explored the sensitivity of GIR to different spatial representations. A range of representations can be used for spatial domain in metadata, from a full polygon to a centroid or minimum bounding rectangle. This paper examined the sensitivity of matching and ranking approaches to different representations and explored the use of a probabilistic ranking method. Finally, the addition of geographic

context in the form of a *shorefactor*, describing the percentage of a query region on or offshore, was explored.

The final paper in this session, *Spatial search, ranking and interoperability*, from Janée and Frew was the first of two papers describing the extensive experiences of the Alexandria Digital Library (ADL) project in GIR. The paper set out the challenges in providing a set of interoperable interfaces sufficiently detailed to allow useful spatial querying and yet sufficiently lightweight to be taken up by many organisations. Key concepts in such work are the nature of spatial representations and the spatial predicates (e.g. overlaps, inside, etc.) which are considered a minimal set required for meaningful spatial queries to be performed. Important issues in this paper included the need for geodetic continuity in representation across the globe and the suggestion that the basic set of spatial predicates used be limited to the relation *intersects*.

The third session, ***Extracting Geographic Information from Unstructured Text***, opened with a paper from Silva et al. titled *Adding geographic scope to web resources* investigated the problem of determining the geographical scope of web documents, in order to index them in the tumba! web search engine. After transforming web documents to a structured XML/RDF format they were progressively augmented with geographical descriptors through a sequence of lexical analysis, geographical entity recognition and semantic and web inference procedures.

The specific aspect of geographical scope determination concerned with recognising the presence of place names in documents was the subject of the paper by Nissim et al., *Recognising geographical entities in Scottish historical documents*. They showed that an existing machine learning tool, specifically a maximum entropy tagger, can be used quite effectively to detect the presence of place names using a manually generated training data set. They worked with a challengingly "noisy" data source, the Statistical Accounts of Scotland.

Amitay et al. presented a paper on *Finding the geographical focus of web-pages* which used the results of a geographic name tagging procedure to try to determine the main geographic focus of a web document when several place names are mentioned. If there are several places mentioned along with their parent region, then the immediate parent is assumed to be a focus of the document. The procedure depends upon the use of a hierarchically-structured gazetteer.

The final paper in this session, *Extracting spatial information: grounding, classifying and linking spatial expressions* looked at the uncertainty introduced into the classification of documents according to the occurrence of place names due to the various uses of place descriptors in text. Schilder et al. used a part-of-speech tagger to identify spatial expressions which were then disambiguated with a search procedure to distinguish between place names used to identify a location and places referring, for example, to governments, to the nationalities of people and to the destinations of goods.

The final session of the workshop, ***Test collections*** provided an opportunity for the participants to discuss possibilities for the development of test collections to be used across the GIR community in a similar way to the successful TREC collections. The first paper in this session, from Hill et al., *Research directions in georeferenced IR*

based on the Alexandria Digital Library Project, provided a further insight into the work of the ADL, and in particular experiences related to georeferencing through the use of gazetteers. The paper proposed the development of a spatial test collection – that is a set of detailed data with which results from the sort of experiments described by Larson and Frontiera could be compared. Such a reference set would provide a consistent reference dataset for measuring spatial relevance in terms of different predicates and representations.

The second paper in this section, *Towards a reference corpus for automatic toponym resolution evaluation* by Leidner, presented a detailed proposal for a reference corpus of text annotated with spatial locations. A detailed discussion of the issues in developing such a corpus was presented, including a review of issues such as the selection of an appropriate corpus, selection of a gazetteer and a proposed mark-up scheme.

This paper was complimented by the final paper of the workshop, from Clough and Sanderson entitled *A proposal for comparative evaluation of automatic annotation for geo-referenced documents.* This paper presented a further proposal for the development of test datasets containing annotated geographic entities. The session closed with a discussion on possibilities for future collaborations on test collections. A very positive result of these discussions was a contribution to CLEF (Cross-language evaluation forum: clef.iei.pi.cnr.it:2002/) discussing the possibilities for establishing a GIR track as part of CLEF.

The workshop closed with a discussion as to its worth and future affiliation. All of those present were happy that the workshop had proved fruitful, and interested in attending future workshops. It was agreed that SIGIR had provided an ideal venue to bring together researchers in GIR, but this should not preclude holding the workshop with other conferences, in line with its interdisciplinary status.

The workshop proceedings are available online at http://www.geo.unizh.ch/~rsp/gir/. A special issue of Computers, Environment and Urban Systems is also planned.