Spatial Information Retrieval and Geographical Ontologies

An Overview of the SPIRIT Project

Christopher B. Jones

Cardiff University, Department of Computer Science, Cardiff, UK, +44 29 2087 4796

R. Purves[1], A. Ruas[2], M. Sanderson[3], M. Sester[4], M. van Kreveld[5], R. Weibel[1]

[1]University of Zurich, [2]IGN Paris, [3]University of Sheffield, [4]University of Hannover,

[5]Utrecht University

Technical Report CS-UU-2002-043

Institute of Information and Computing Sciences

Utrecht University

Spatial Information Retrieval and Geographical Ontologies An Overview of the SPIRIT Project

Christopher B. Jones
Cardiff University, Department of Computer Science, Cardiff, UK, +44 29 2087 4796 {c.b.jones@cs.cf.ac.uk}
R. Purves[1], A. Ruas[2], M. Sanderson[3], M. Sester[4], M. van Kreveld[5], R. Weibel[1]
[1]University of Zurich, [2]IGN Paris, [3]University of Sheffield, [4]University of Hannover, [5]Utrecht University

Categories & Subject Descriptors: H.3.3 [Information Storage and retrieval]: Information Search and Retrieval - information filtering, query formulation, relevance feedback, search process.

General Terms: design.

1. INTRODUCTION

A large proportion of the resources available on the world-wide web refer to information that may be regarded as geographically located. Thus most activities and enterprises take place in one or more places on the Earth's surface and there is a wealth of survey data, images, maps and reports that relate to specific places or regions. Despite the prevalence of geographical context, existing web search facilities are poorly adapted to help people find information that relates to a particular location. When the name of a place is typed into a typical search engine, web pages that include that name in their text will be retrieved, but it is likely that many resources that are also associated with the place may not be retrieved. Thus resources relating to places that are inside the specified place may not be found, nor may be places that are nearby or that are equivalent but referred to by another name. Specification of geographical context frequently requires the use of spatial relationships concerning distance or containment for example, yet such terminology cannot be understood by existing search engines.

Here we provide a brief survey of existing facilities for geographical information retrieval on the web, before describing a set of tools and techniques that are being developed in the project SPIRIT : Spatially-Aware Information Retrieval on the Internet (funded by European Commission Framework V Project IST-2001-35047).

1.1 Current Support for Geographical Information Access on the Web

The Northern Light GeoSearch tool (http://www.northernlight.com/geosearch.html) from Vicinity includes a facility for geographical information retrieval. It allows the user to enter part or all of an address, along with a category of interest and a search radius. It finds other places within the specified radius, with the aid of a digital map. A similar facility is found on www.somewherenear.com. Retrieval facilities to find resources associated with a specific place or post code (or zip code) are to be found in association with yellow pages services and some digital map web sites.

A major shortcoming of these approaches is that they cannot recognise alternative names for the same place, whether they be modern and historical variants, informal names or names of contained places. Furthermore their user interfaces are limited in the options they offer to specify and refine a search and to visualise the results, reflecting their lack of knowledge of the semantics of geographical terminology and the associated spatial structures. There have been several recent initiatives to develop location-specific referencing of internet domain names.

Buyukokkten [1] associated IP addresses with telephone area codes of the associated network administrators, and hence, via zip code databases, to place names and geographical coordinates. The approach facilitates the analysis of the geographical distribution of web sites, but is limited in retrieving web pages on the basis of subject matter and geography, since it cannot be assumed that the content of a web page is related to the place where it was created.

The Stanford Research Institute proposed a top level domain based on geographical referencing (.geo). The domain name refers to a strict hierarchy of quadrilateral cells defined by latitude and longitude. Existing domain names would be able to register themselves with a *.geo* domain server which would store, for a set of cells, all registered web sites that relate to each of the given cells. Effective geographical search would depend upon the ability to map from the user's specification of geographical location to the relevant cells. Thus the approach still leaves the requirement for a semantic level of retrieval tools to exploit it.

An experimental system for geographical navigation of the web has been described by [2]. Techniques are proposed for extraction of the geographical context of a web page, based on the occurrence of text addresses and post codes, place names and telephone numbers. This information is transformed to one of a limited set of point-referenced map locations. Geographic search is initiated by the user asking to find web sites that refer to places in the vicinity of a currently displayed web site. The geographical user interface does not allow user specification of place names.

3. TECHNIQUES FOR SPATIALLY-AWARE INFORMATION RETRIEVAL

The SPIRIT project addresses shortcomings of existing facilities such as those described by developing:

- ontologies that model geographical terminology;
- query expansion and relevance ranking procedures based on the geographical ontologies;
- machine learning techniques for the extraction of geographical context from web documents and for generating metadata providing spatial context;
- a multi-modal user interface providing textual input and interactive map feedback of the context of retrieved documents;
- spatial indices for web collections.

These facilities are to be tested and evaluated with a 1Tb web collection that will be stored and indexed within a parallel architecture.

2.1 Geographical and conceptual ontologies

The geographical ontology within SPIRIT models both the vocabulary and the spatial structure of places for purposes of information retrieval. The approach facilitates the retrieval of resources that employ alternative and spatially-related place names to that in the query. This will result in accessing not just those resources with an exact locational match to the terms of the query, but also those that relate to places that are similar or nearby in location to the query. The search facilities will furthermore recognise different versions of place names as well as historical alternatives. The potential of building geographical models that combine qualitative and quantitative spatial information for purposes of information retrieval has been recognised in literature on gazetteers and geographical thesauri [3] [4] [5], but has yet to be proven. This project extends the concepts of gazetteers and thesauri to include a richer set of qualitative and quantitative spatial properties that can be exploited for information retrieval.

In the field of IR, automatic query expansion to add terms to the user's query has received much attention. Although techniques using statistically derived associations have proven useful [6], methods using thesauri with synonyms have shown less success [7]. This has been due to ambiguity of the query terms and its propagation to synonyms that may be unrelated to the user's interests. In geographical terminology, ambiguities are relatively easily resolved and there is good reason to believe that, with very simple user dialog, effective term expansion can be achieved.

2.2 Geographical relevance ranking

The geographical ontology maintains qualitative and quantitative information that can be used to create geographical similarity measures and document ranking techniques using methods analogous to [5]. In the fields of image recognition and general pattern matching, geometric similarity measures have been a topic of study for many years [8]. In parallel, semantic similarity measures have been developed in the field of information retrieval, e.g. [9]. We draw upon both of these strands to develop novel similarity measures, combining geometric, qualitative spatial, and non-spatial information that will be evaluated with large web data collections.

2.3 A multi-modal user interface for geographical information retrieval

Multi-modal interfaces are known to be an effective way to interact with geographical information, yet no research has been conducted in this area for geographical information retrieval. SPIRIT users will be able to describe geographical location textually with terminology for place names, spatial relationships and information descriptors. The ontologies will help to disambiguate the query and to reflect its geographical extent textually and on a map, revealing the system's interpretation of imprecise qualitative spatial terms relating to direction and proximity. Alternatively the user will be able to mark the required extent of search on the map. This could include selecting displayed place names, drawing regions of interest, and using displayed map features to delineate the region of interest. The level of detail of the map will adapt to the user's interests and will reflect the hierarchical structures of geographic space encoded in the ontology.

2.4 Geographical metadata enrichment

Intelligent use of information on the internet depends upon it being accessible to mobile agents and search engines. This depends in turn upon essential aspects of the content of the documents being abstracted and made available in the form of semantic annotations. The process of annotation however requires methods of extracting the information.

In the context of spatial data annotation, progress has been made in the production of metadata *encoding* methods that describe the location and quality of the data sets [10]. The process of annotation is however typically manual. Here we develop techniques for *automatic extraction* and metadata encoding of information from digital map datasets, exploiting techniques from computational geometry, image interpretation and spatial data mining. We will also develop geographical metadata extraction techniques for unstructured web documents, employing a machine learning approach, using training datasets of documents that have already been accurately tagged on the basis of key words such as place names, addresses, telephone numbers and postcodes.

4. REFERENCES

[1] Buyukokkten, O., et al. Exploiting geographical location information of web pages. in WebDB'99 (with ACM SIGMOD'99), 1999, 91-96.

[2] McCurley, K.S. Geospatial mapping and navigation on the web. in WWW10, 2001 http://www10.org/cdrom/papers/278/

[3] Harpring, P., Proper words in proper places: The Thesaurus of Geographic Names. MDA Information, 1997, 2(3): 5-12.

[4] Hill, L.L., J. Frew, and Q. Zheng, Geographic Names. The implementation of a gazetteer in a georeferenced digital library. Digital Library, 1999, 5(1): www.dlib.org/dlib/january99/hill/01hill.html.

[5] Jones, C.B., H. Alani, and D. Tudhope Geographical information retrieval with ontologies of place, in Spatial Information Theory, LNCS 2205, 2001 322-35.

[6] Xu J and W.B.Croft Query Expansion Using Local and Global Document Analysis.ACM SIGIR 1996 4-11.

[7] Voorhees, E.M. Query expansion using lexical-semantic relations. in ACM SIGIR 1994: 61-70.

[8] Veltkamp, R.C. and M. Hagedoorn State-of-the-art in shape matching, in Principles of Visual Information Retrieval, M. Lew, Editor, 2001, Springer, 87-119.

[9] Sintichakis, M. and P. Constantopoulos. A method for monolingual thesauri merging. in ACM SIGIR 1997, 129-138.

[10] WMS, OpenGIS® Web Map Server Interfaces Implementation Specification, 2001, http://www.opengis.org/techno/specs.htm.