Spatial Information Retrieval

Wenwen LI[1, 2], Phil Yang[1], Bin Zhou[1, 3]
[1] Joint Center for Intelligent Spatial Computing, and Earth System & GeoInformation Sciences College of Science, George Mason University, U.S.
[2] Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China
[3] Department of Electronic Engineering, Tsinghua University, Beijing, China

SYNONYMS

Information Retrieval (IR), Geographical Information Retrieval (GIR)

DEFINITION

Spatial Information Retrieval (SIR) refers to providing access to geo-referenced sources by indexing, searching, retrieving, and browsing [1]. SIR is an interdisciplinary topic involving geospatial information science, data mining, computer network, cognition, and cartography. An example of SIR query could be “Where are the Italian restaurants within 500 meters around a lake park?” Figure 1 (a) illustrates the spatial distribution around the park: Shaded area represents the park; triangles represent restaurants; the red triangles represent those providing Italian food. To answer the question, spatial entities, such as restaurants and the park, are abstracted to a point described in [x, y] coordinates. To execute the query, both Information Retrieval algorithm and spatial relationship analysis are needed: 1) locate the area of interest where a circle has a radium of 500 meters (Figure 1b); 2) filter out restaurants from other objects (represented by the point in Figure1); 3) search the restaurant index to identify the keyword 'restaurant types' with 'Italian food'; 4) a list of restaurants, ranked by the distance to the park, is sent back to users.

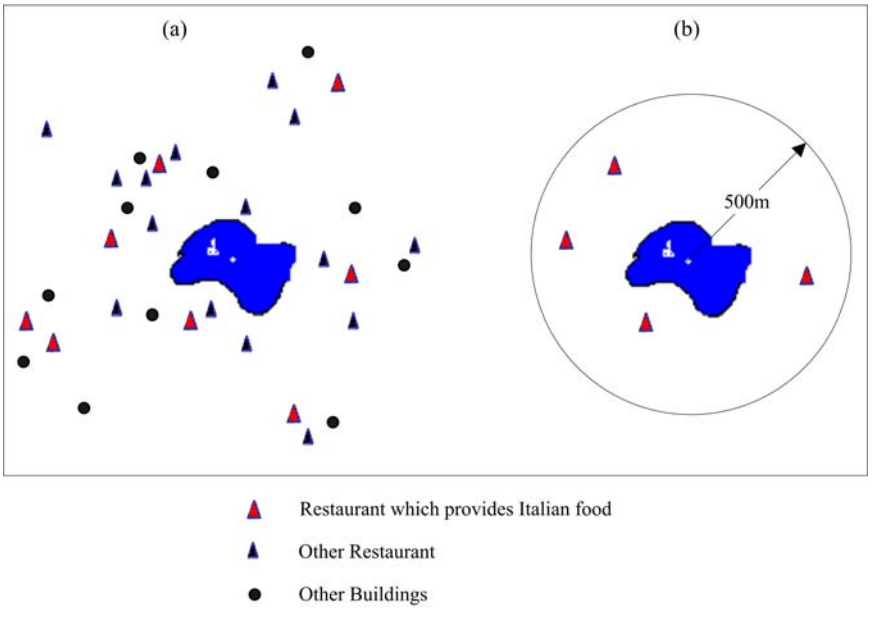

Figure 1. A Query example of Spatial Information Retrieval

SPATIAL DATA SOURCES

Spatial retrieval procedures are different from each other according to different types of spatial data sources, such as Digital Library and World Wide Web.

SIR IN DIGITAL LIBRARY

In general, Digital Library stores large volume of categorized information, for example, water resources below ground in a valley for flood analyzing and monitoring. Researchers [18] proposed automatic geospatial information processing system to provide fast access to Digital Library. In such a system, Geographic place names (terms) and their attributes are extracted and identified based on thesaurus and semantic information containing spatial relationships, such as “adjacent to a lake”, “south of the river”. Geographical coordinates are retrieved and probabilistic weights are assigned to the place names based on their occurrence in thesaurus. Therefore, each term can be denoted as a 3-D object, dimensions [x, y] represent geographic coverage, and dimension Z represents weight of the term. Finally, all the terms extracted are denoted in 3-D space to form a ‘skyline’, where weights are summed when two terms have overlaps in geographical coverage. The geographic area where the peak of the ‘skyline’ located will be indexed. Then, by applying the algorithm to all texts stored in the Digital Library, the entire index can be established to assist fast access.

SIR IN WORLD WIDE WEB

The context is complex in Internet-based spatial information retrieval because World Wide Web (WWW) contains huge amount of information. For example, as estimated by Danny Sullivan in 2005, Google indexed more than 9 billion documents in its crawler, where the web documents collection is built and maintained through crawling. Meanwhile, performance stood out as an issue for spatial indexing in such a large collection. To support spatial indexing in large amount of documents, researchers [8] proposed Spatio-textual indexing algorithm with the help of geographical ontology. Each document containing place names is associated with one or more ‘footprints’ (using coordinates to present a place) derived from ontology entries [8]. Then a three-step algorithm is applied: 1) all the documents are divided to several cells (S_i, i=1, 2…m) based on their spatial distribution marked in the footprint; 2) the document sets (D_j, j=1, 2… n) indexed by key words are intersected with each space cell (S_i) to form the spatio-textual index (S_i, D_j), Therefore, an index with this structure can be exploited by first searching for a textual term; then 3) the associated spatial index of documents are used to filter out those meeting the spatial constraints [8] so that the ambiguity of terms can be eliminated and query accuracy and performance is improved.

HISTORICAL BACKGROUND

The history of SIR can be traced back to year 850, when the first print book was created in China that changed the traditional mechanism of information storage. The second leap was in 1946, when the first electronic computer transformed data into digital version. In 1950, *Information*

Retrieval was first used by Vannevar Bush, and became popular [2]. In 1996, Ray R. Larson coined *Spatial Information Retrieval* in Digital Library's service [1]. Many SIR models have been designed and applied to retrieve geospatial information [3 and 12].

SCIENTIFIC FUNDAMENTALS

In general, an SIR system deals with spatial queries such as "what is here?" asking for place names, geographical features about a location or "where is it?" resulting in a reference in a map [4]. Five types of spatial queries are summarized as [1]: *Point-in-polygon query*, which is a relatively precise query asking about the geographic information of a point (denoted by [x, y]) within a area (denoted by polygon); *Region query*, asking for any geographical element that is contained in, adjacent to or overlaps the region defined; *Distance and Buffer Zone query* (Figure 1) refers to finding spatial objects that are within certain distance of an area; *Path Query*, which requires the presence of a network structure to do shortest path or shortest time planning and *Multimedia query*, which combines both geo-reference processing and non-georeference processing, such as pattern recognition, in executing a query.

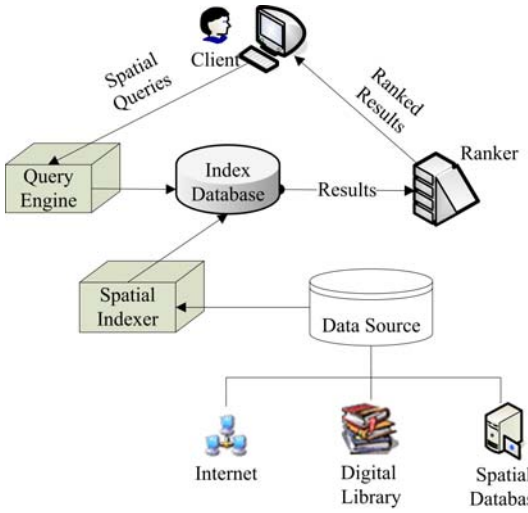

Figure 2. A General Model for Spatial Information Retrieval

A general SIR model for answering previous spatial queries includes data source, spatial indexer, query engine, and ranker (Figure 2). Spatial information is obtained from geospatial data. The data are in different formats, such as textual, numerical, graphical, and multimedia. Spatial indexer provides spatial indexing by extracting geographic locations in a text or mapping data to a term based on certain geospatial ontology [6, 7]. Query engine handles user requests [5]. To provide a better Quality of Service, a ranker is normally used to sort the results based on match level.

Geospatial information sources for spatial retrieval are generally available from Digital Library or WWW, where large amount of data are stored in a variety of formats including basic text documents, airborne and satellite images, maps from specific geographic locations, and other

forms. Therefore, it's of great importance to extract and index the spatial element from data sources to improve query efficiency [9, 10]. Meanwhile, the dynamic, incoherent nature of WWW makes it more difficult for SIR system to gather information, make spatial indexing structures scalable and efficiently updatable. The solutions to these problems relying on probability theory and spatial reasoning as discussed in Key Applications.

KEY APPLICATIONS

Spatial Information Retrieval is widely used in applications ranging from scientific research, government planning, to our daily life.

EARTH AND PLANETARY RESEARCH

Terabytes of imagery data are collected through satellite and airborne remote sensors every day [11].While providing valuable resource to earth system research, the imagery also complicates the management of the data. SIR could help to get the data with needed appropriate spatial characteristics, time span, and appropriate format from vast amount of datasets. Project Sequoia [12] gives a successful example for earth scientists to retrieve information from tens of terabytes of spatial and geographical data.

DISASTER AND DISEASE ANALYSIS

SIR can assist researchers and policy makers to extract emergency information, for example, the asset loss of cities during a river flooding [13, 14].

URBAN TRANSPORTATION

Spatial Information Retrieval can be applied to urban transportation management by indexing and analyzing the transportation datasets.

ENVIRONMENT PROTECTION

U.S. Environment Protection Agency (EPA) has its Aerometric Information Retrieval System and national Pesticide Information Retrieval System for viewing and researching the spatial distribution and trends for pollution or other environment destructions [15].

TRAVELING

SIR can be integrated in mobile device to help travelers in guidance and travel routes planning [1, 16].

FUTURE DIRECTIONS

Any query related to location needs the support of SIR system. Future SIR System will be more intelligent and be able to answer questions in natural language description. Meanwhile, it'll play a more important role in contemporary geospatial processing, especially in the context of WWW.

CROSS REFERENCE

Spatial Indexing, Spatial Reasoning, Spatial Ontology, Semantics, and Taxonomy

RECOMMENDED READING

[1] R.R. Larson. Geographical Information Retrieval and Spatial Browsing. In *In Geographical Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pages 81-124, 1996.

[2] http://online.sfsu.edu/~fielden/hist.htm

[3] C. B. Jones. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project. In *In the work SIGIR'02*, 2002.

[4] Oyvind, V. Geographic Information Retrieval: An Overview

[5] A. Arasu, J. CHO, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. In *In ACM Transaction on Internet Technology*, volume 1, pages 2-43, 2001.

[6] K.S. McCurley. Geospatial Mapping and Navigation of the Web. In *In the Proceedings of the 10th international conference on World Wide Web*, pages 221-229, 2001.

[7] A.I. Abdelmoty, P.D. Smart, C.B. Jones, G.. Fu and D. Finch. A Critical Evaluation of Ontology Languages for Geographical Information Retrieval. In *In the Journal of Visual Languages and Computing*, volume 16, pages 331-358, 2005.

[8] C. B. Jones, A. I. Abdelmoty, D. Finch, G.. Fu and S. Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *In Proc Third International Conference on Geographic Information Science GIScience, Lecture Notes in Computer Science 3234*, pages 125-139, 2004.

[9] O. Buyukokkten, et al. Exploiting Geographical Location Information of Web Pages. In *In WebDB'99 (with ACM SIGMOD'99)*, pages 91-96, 1999.

[10] D. Mountain and A. MacFarlane. Geographic Information Retrieval in a Mobile Environment: Evaluating the Needs of Mobile Individuals. To be published In *In the Journal of Information Science*, 2006.

[11] C.W. Emerson, D.A. Quattrochi and N.S. Lam. Spatial Metadata for Remote Sensing Imagery. *In In the 4th annual Earth Science Technology Conference (ESTC)*, 2004.

[12] J. Chen, R. R. Larson and M. Stonebraker. Sequoia 2000 object browser. In *Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference*, pages 389-394, 1992.

[13] http://www.em-dat.net/links/disasterdbs.html

[14] W. Lu, S. Mannen, M. Sakamoto, O. Uchida and T. Doihara. Integration of Imageries in GIS for Disaster Prevention Support System. In *In XXth ISPRS Congress*, 2004.

[15] http://www.epa.gov/enviro/html/airs/index.html

[16] P. Clough. Extracting Metadata for Spatially-aware Information Retrieval on the Internet. In *In the Proceedings of the 2005 workshop on Geographic information retrieval*, 2005.

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

[18] A.G. Woodruff, C. Plaunt. GIPSY: Geo-referenced Information Processing System. In *Journal of the American Society for Information Science*, volume 45(9), pp 645-655, 1994.