A Hybrid Indexing and Ranking Approach to Enhance Geospatial Semantic Search

Wenwen Li

GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, 975 S. Myrtle Ave., Tempe, AZ 85287-5302
Telephone: 1-480-727-5987
Fax: 1-480-965-8313
Email: wenwen@asu.edu

1. Introduction

Nowadays, geospatial information has been extensively used to support a variety of physical-science and social-science studies, such as natural-disaster prediction (Li et al. 2009), emergency response (Rauschert et al 2002), and urban-economics studies (Anas and Liu, 2007). In the past decades, billions of gigabytes of geospatial data have been produced and made available to the public by government agencies and other stakeholders from multiple Earth-orbit missions, ground survey, and in-situ measurements. The large volume of data provides science and applied researchers with a valuable resource. To enable the seamless access and visualization of geo-referenced data, the former Vice President of the United States Al Gore envisaged a virtual globe -- the Digital Earth -- as "a new wave of technological innovation that allows us to capture, store, process and display an unprecedented amount of information about our planet and a wide variety of environmental and cultural phenomena" (Gore 1998). Ten years later, a number of advanced techniques, such as geobrowsing, distributed geographic information processing (DGIP, Yang et al. 2008), and volunteered geographic information (VGI; Goodchild, 2007), have been developed to operationalize the Digital Earth concept. However, as a comprehensive goal, the "Digital Earth" is still facing challenging problems (Xu, 1999; Craglia et al 2008). One grand challenge is how to provide an intelligent mechanism to assist users of Digital Earth systems to readily discover, search, and access useful science content from multiple sources. In the position paper from the Vespucci Initiative for the Advancement of Geographic Information Science, Craglia et al. (2008) highlighted the importance of establishing "a dynamic information system to provide reliable, accurate, timely and openly accessible information" for building the next generation Digital Earth. In 2010, the workshop "Towards Digital Earth: Search, Discover and Share Geospatial Data 2010" (http://ceur-ws.org/Vol-640/) was held at the Future Internet Symposium and discussed the application of state-of-the-art information technology to enable intelligent discovery of geospatial data. Although several efforts have been made to promote the scientific discovery process, such as establishing data application centers and developing Web catalogs (Liu et al. 2010) with search capabilities, in reality, scientists are still limited to the use of datasets that are familiar to them (Li et al. 2011). These efforts often have little knowledge of the existence of datasets that could be a better fit for their model or application (Gray et al 2005; Singh 2010; Tisthammer 2010) due to the inefficiency of current geospatial search engines. This deficiency brings great challenges to the information-retrieval community to develop more effective mechanisms for intelligent geospatial data discovery and a semantic-search platform to support the realization of the Digital Earth vision (Gore 1998; Li et al 2008a; Li et al. 2008b).

There are two factors that influence the discoverability of a geospatial search engine in the digitized world: accessibility and effectiveness. Accessibility measures whether all existing geospatial data and services can be accessed by as many users as possible; in other words, it

involves the process of building the corpus which provides the most up-to-date data. Effectiveness measures whether a search engine is able to find all relevant information by scanning the corpus. One way to improve accessibility is to build a comprehensive corpus containing all available datasets dispersed on the Internet. For example, NASA has built several distributed, discipline-specific active archive centers (DAACs) for scientific modeling and analysis. NASA's Global Change Master Directory (GCMD) and the US Government's Geospatial One Stop (GOS) provide public gateways and catalogs to facilitate the collection and access of geospatial data. Li et al. (2010) developed an active crawler to automatically collect the distributed geospatial services that exist on the Web and have not yet been published, and to incorporate them into the above catalogs to extend the geospatial data corpus. These works have greatly improved the accessibility of geospatial data. However, in terms of improving the effectiveness of a search engine, almost all of the existing geospatial catalogs and Web portals use Lucene, a full-text keyword-matching technique (Hatcher and Gospodnetic 2004). The datasets that are semantically related to a user's query but described differently from the query keyword will be considered irrelevant and excluded from the search results. Hence, improving the effectiveness of a geospatial search engine and making available datasets reachable by scientists is becoming even more significant.

Recently, the emerging semantic technologies are attracting the attention of researchers, who are exploring how to utilize such technology to improve search effectiveness. One direction of the efforts is to incorporate domain ontologies to identify associations and concepts (such as polyseme, synonym) related to a query, recommending a list of related search terms for users to refine their search. These works include VSTO (Fox et al. 2008), GEON (Bowers et al. 2004), LEAD (Droegemeler et al. 2005) and Noesis (Movva et al. 2008). These solutions rely heavily on the logical representation in the ontology, which is usually developed by humans. The issue is that the words used for indexing a document are often different from those in the pre-defined ontologies. Moreover, different people with different knowledge sets tend to have different perspectives on the categorization of terms and their linkages and relations. This would cause heterogeneous representations and conflicting statements, and eventually influence the effectiveness of a search engine. To overcome this problem, in this paper we propose to use an analytical and human-independent method -- Latent Semantic Analysis (Dumais 2004) -- which has rarely been applied to the retrieval of geographic data. By applying latent semantic analysis, the semantic structure of documents in the corpus can be discovered and the latent semantics between the occurrences of patterns of words, and clues to the likely occurrence of others, will also be discovered. In this way, even the words with no occurrence in a document will be given weights indicating the correlation between the words and the document.

Latent semantic analysis (LSA) enables the discovery of more semantically relevant datasets. Meanwhile, these discovered dataset need to be ranked so that the most relevant results will always appear on top. Therefore, we also propose a ranking model based on revised cosine similarity to filter out documents that are not closely related in order to improve the effectiveness of geographic data retrieval. The geospatial metadata sets from the NASA SEDAC (Socio-Economic Data Application Center) are used as our test corpus in this study. Experiments show that a retrieval system implementing the proposed method improved the retrieval of relevant documents significantly – for all eight sample subject-based queries, the recall rate almost reached 100%. Although the precision is in some cases lower than the Lucene-based retrieval

method, the system guarantees that all the records returned by Lucene could be discovered by the proposed retrieval system. Besides the capability of handling subject-based queries, we also introduced the advanced mechanisms of automatic place-name detection and spatial filtering to handle spatial queries with the assistance of the GCMD location taxonomy.

2. References

Anas, A. and Liu, Y. 2007. A regional economy, land use, and transportation model (relu-tran): Formulation, algorithm design, and testing. *Journal of Regional Science*, 47(3), 415-455.

Bowers, S., Lin, K. and Ludascher, B., 2004. *On Integrating Scientific Resources through Semantic Registration*. Proceedings of the 16th International Conference on Scientific and Statistical Database Management. IEEE Computer Society, 349.

Craglia, M., Goodchild, M.F., Annoni A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S. and Parsons, E. 2008. Next-Generation Digital Earth – A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. *International Journal of Spatial Data Infrastructures Research*, 3, 146 –167.

Deerwester, S., *et al.* 1990. *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6), 391-407.

Droegemeier, K., *et al.* 2005. *Service-oriented environments for dynamically, interacting with mesoscale weather*. Computing in Science & Engineering, 7(6), 12-29.

Dumais, S. T. 2004. *Latent semantic analysis*. Annual Review of Information Science and Technology, 38, 189-230.

Fox, P., et al. 2009. *Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience*. Computers and Geosciences, 35(4), 724-738.

Gabrilovich, E. and Markovitch, S. 2009. *Wikipedia-based Semantic Interpretation for Natural Language Processing*. Journal of Artificial Intelligence Research, 34, 443-498.

GEO, 2005. *The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan* [online]. Available from: http://earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf (Accessed 8 December 2011).

Goodchild, M.F., 2007, *Citizens as sensors: the world of volunteered geography*, GeoJournal 69(4): 211-221.

Gore, A., 1998, *The Digital Earth: Understanding our planet in the 21st Century, Given at the California Science Center*. Los Angeles, California on Jan. 31, 1998.

Gray, J., Liu, D. T. and Dewitt, D. J. 2005. *Scientific data-managernent in the coming decade*. Sigmod Record, 34(4), 34-41.

Hatcher, E. and Gospodnetic, O., 2004. *Lucene in Action*. Greenwich, CT, USA: Manning Publications Co.

Li, W., Yang, C., Nebert, D., Raskin, R., Houser, P., Wu, H., Li, Z., 2011. *Semantic-based service chaining for building a virtual Arctic spatial data infrastructure*. Computers & Geosciences, 37(11), 1752-1762.

Li, W., Yang, C. W. and Yang, C. J. 2010. *An active crawler for discovering geospatial Web services and their distribution pattern - A case study of OGC Web Map Service*. International Journal of Geographical Information Science, 24(8), 1127-1147.

Li, W., Yang, C. and Sun, D. 2009. *Mining geophysical parameters through decision-tree analysis to determine correlation with tropical cyclone development*. Computers & Geosciences, 35(2), 309-316.

Li, W., Yang, C. and Raskin, R., 2008a. *A semantic enhanced model for searching in spatial web portals*. In: Proceedings of Semantic Scientific Knowledge Integration AAAI/SSKI Symposium, Stanford U., Palo Alto, CA, US, 47-50.

Li, W., Yang, C. and Zhou, B. 2008b. *Internet-based spatial information retrieval*. Encyclopedia of GIS, 1, 596-599.

Movva, S., *et al.*, 2008. *Customizable Search Engine with Semantic and Resource Aggregation Capability*. Proceedings of the 2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services. IEEE Computer Society, 376-381.

Park, L. A. F. and Ramamohanarao, K. 2009. *Efficient storage and retrieval of probabilistic latent semantic information for information retrieval*. Vldb Journal, 18(1), 141-155.

Rauschert, I., *et al.*, 2002. *Designing a human-centered, multimodal GIS interface to support emergency management*. Proceedings of the 10th ACM international symposium on Advances in geographic information systems %@ 1-58113-591-2. McLean, Virginia, USA: ACM, 119-124.

Singh, D. 2010. *The biological data scientist. Business, Bytes, Genes and Molecules*.
TAN, P.-N., STEINBACH, M. and KUMAR, V., 2006. Introduction to data mining. 1st ed. Boston: Pearson Addison Wesley.
Tisthammer, W. A. 2010. *The nature and philosophy of Science*. UFO Evidence, 1386.
Xu, G. 1999. *Meeting the challenge of "Digital Earth"*. Journal of Remote Sensing, 3(2): 85-89.
Yang, C., LI, W., Xie, J. and Zhou, B. 2008. *Distributed geospatial information processing - sharing distributed geospatial resources to support the Digital Earth*. International Journal of Digital Earth, 1(3), 259-278.