The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-1/W5, 2015
International Conference on Sensors & Models in Remote Sensing & Photogrammetry, 23–25 Nov 2015, Kish Island, Iran

TAGS EXTARCTION FROM SPATIAL DOCUMENTS IN SEARCH ENGINES

S. Borhaninejad [a], F. Hakimpour*, E. Hamzei

[a] School of Surveying and Geospatial Engineering, Faculty of Engineering, University of Tehran
(saeid.borhani, fhakimpour, e.hamzei)@ut.ac.ir

KEY WORDS: spatial; search engine; GML documents; crawler; spatial documents; data access

ABSTRACT:

Nowadays the selective access to information on the Web is provided by search engines, but in the cases which the data includes spatial information the search task becomes more complex and search engines require special capabilities. The purpose of this study is to extract the information which lies in spatial documents. To that end, we implement and evaluate information extraction from GML documents and a retrieval method in an integrated approach. Our proposed system consists of three components: crawler, database and user interface. In crawler component, GML documents are discovered and their text is parsed for information extraction; storage. The database component is responsible for indexing of information which is collected by crawlers. Finally the user interface component provides the interaction between system and user. We have implemented this system as a pilot system on an Application Server as a simulation of Web. Our system as a spatial search engine provided searching capability throughout the GML documents and thus an important step to improve the efficiency of search engines has been taken.

1. INTRODUCTION

The freedom of information concept requires that all those who somehow can use the information, access that information, So that with the least knowledge about data set, selective data recovery is possible (Jansen and Spink, 2006). In general selective access to information on the Web is provided through the search engines, but in the cases that our queries includes spatial information, searching is more complicated and requires special capabilities in the search engine system.

Nowadays arrival of spatial information on the Web makes spatial search engines drift from text and multimedia processing to spatial information processing.

All aspects of human activity are rooted in geographical space. As a result, many of the documents include geographical texts references that are usually location names (Jones *et.al*, 2004). This is a common occurrence in the World Wide Web documents (World Wide Web). If a user wants the search engine to find the sources that are related to a specific location, the user can place his desired location name in the search engine query. The behaviour of conventional search engines toward location names (words with spatial semantics) is the same as other keywords and retrieved documents containing that specified names (Purves, 2007). For some purposes it may be enough, but there are many situations in which the user is interested in documents that relate to the same region of space as that specified by the place name, but which might not actually include the place name. So we require a location-aware search engine (spatially-aware) that can intelligently interpret the location names in query so that high quality results can be retrieved.

A document which contains a location on the Web can include spatial information such as features, coordinates, spatial relationships and spatial descriptions. Text processing of such data, and extracting spatial information from it can be very effective on results presentation and retrieval.

A lot of research has been done on the HTML documents text processing, but spatial texts like GML processing remains a challenging topic. Despite the fact that today's engineers and specialists in many fields need raw spatial data and looking for it on the World Wide Web, most of spatial search engines are based on map representation and less attention is paid to spatial data.

Our research, focuses on the recovery of data from the spatial documents and specifically GML documents, so by creating an integrated attitude, we can extract the spatial and non-spatial information from these documents. General attitude to retrieve spatial information is in a way that, spatial information can be extracted through their connection to non-spatial information, While the spatial and non-spatial information seamlessly stored in the documents. One approach is provided in (Zhao, 2013). They provide a framework for the use of the position information to search the web. In fact, the authors aim was spatial query processing to access web pages position information. In this approach, despite considering the position information of web pages which traditional search engines pays less attention to that, the spatial attitude is the same general attitude and the spatial information retrieved by their connection to the non-spatial information.

Similar to Jie Zhao and et al, a lot of works done to explore the web pages position. (Wang, 2005) proposed an algorithm for extracting the Web queries positions. Wang and his co-authors classified the position of Web sources into three categories include: provider location, content location and serving location. They used hyperlinks, user blogs and Web content to discover these three types of position. (Wang, 2010) proposed a heuristic four step algorithm called web-a-where to determine the focused position of the Web pages, in which, all names are assigned to a position with a degree of confidence. Based on degree of confidence as well as other parameters such as frequency and spatial relationship, the focused position of a web page extracted. One solution to search for spatial data over the web is using a web crawler technology that it can search throughout the web for spatial information (Li, 2010 and Walter, 2013). Crawlers are Web applications which review the Web with the aim of content indexing. The crawlers initially visited the list of specified URL and then by identifying hyperlinks and adding the URL to a frontier list spreads the original list (Bones, 2014). Then the sites in the frontier list visited recursively based on a defined set of rules to determine if it includes the search criteria or not. With this work, crawler

provides a new list of sites that have some criteria, while unrelated and defective URL removed. A study in 2014 has been conducted by (Bones, 2014) to explore the spatial data in various formats. The authors have created a search engine named "GSE" which supports WMS, ArcGIS services and spatial data included web sites. GSE is a crawler based search engine. GSE crawler's seeding mechanism is that the user search terms are combined with the predetermined keywords that identify the spatial data services. However, this approach is dealing directly with spatial documents, but also there is no integrated spatial data retrieval approach, because the GSE focus is on spatial services such as ArcGIS services and WMS and it not care about spatial document's texts. The GSE is looking for a new spatial web services, and it is dismissive about documents texts that include integrated spatial and non-spatial information.

All in all, in this study the following objectives discussed:

1- Extraction of spatial information which is embedded in Web documents: Spatial documents include spatially explicit information such as the coordinates of the feature or the type of feature that extracting this information improves the response rate of spatial queries in search engines.

2- Implementation and evaluation of an integrated spatial information retrieval approach.

2. METHODOLOGY

Despite the fact that today's engineers and specialists in many fields need raw spatial data and looking for it on the World Wide Web, most of spatial search engines are based on map representation and less attention is paid to spatial data. There is a substantial volume of spatial documents and information on the Web, however, the extent of the Web has caused this huge volume of documents and information hard to find among other information. So our proposed system plans to process this type of data and extract the spatial information from it for improving the spatial search engines efficiency. For example, a user who is looking for Tehran and knows only the coordinates of a point in this city, by data retrieval the search engine determines which documents include information in vicinity of that point then an extent of Tehran returns as an output. Similar works only rely on non-spatial and textual documents, and these documents usually do not involve explicit spatial information such as location coordinates or feature type, while spatial document in addition to descriptive information contains explicit spatial information too.

2-1 System overview

Our proposed system consists of three basic components:

1- crawler: The main innovation of this study is this component. The typical spatial search engines crawlers analyse and process mainly HTML documents and extract spatial information contained in these documents. In our proposed system, the crawler processes GML documents text as well as HTML and other textual documents, and extracts the spatial information from these documents. Crawler in this system has two main tasks:

- Detection of GML documents among the documents with different formats.
- Analysis of GML documents and extracting the spatial information

2- Database: database is responsible for storing data which collected by crawlers

3- User Interface: this module provides interaction between user and system and users send their queries through this interface

Searching in this system is possible in two ways: keyword-based method and coordinate-based method.

In searching by keyword, user enters his desired keyword into the system. Since keywords are stored as attributes in the GML documents, system searches for the user desired keyword among the stored attributes of documents, and returns the documents which includes that identical keyword.

In search by coordinates, user enters the coordinates of his desired point into the system. The system calculates the distance between the desired point and GML documents stored features and if the distance was less than a threshold value, the feature included in the document is returned to the user. In polygon features, bounding box of polygon is used instead of distance, in such a way that if the target point is in the bounding box then the document containing that polygon is returned to the user.

In general, this system's search process is done in two phases: online and offline. Offline phase includes the crawler's searching and storing the information in to the database. And the online phase includes user interface and ranking operation.

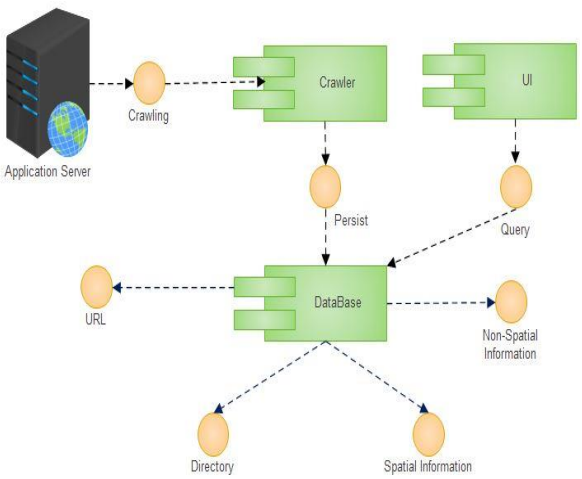

Figure 1. System architecture

2-2 Online Phase

2-2-1 Crawler design

A crawler is embedded in the proposed system to find the GML documents. As mentioned above, clawer's duty is finding and parsing of GML documents. We used Crawler 4J as our proposed system's crawler and with changes in the crawler source code. At this stage of the research our crawler is only sensitive to GML documents. Thus crawler instigates just after receiving the initial seed and starts to find the GML documents and stores its URLs in the database. We implemented our pilot proposed system on a Web Application Server as a simulation of Web space. In this way the pilot version of crawler instead of searching the Web, worked on the data in the application server. As prerequisite of this process, several HTML documents linked with GMLs are provided in the application server. About 10% of this HTML files are inserted as initial seed to crawler.

In the parsing section of GML documents, once crawler find a GML document, it parses text and depending on the nature of each element, inserts the parsed text into separate database tables.

2-2-2 Database design

Database controls documents storage. After the crawler finds GML documents, depending on the nature of each element in the GML texts, various gained information stored in separate tables in the database. Initially crawler stores the URLs of GML documents in a particular table. Then after parsing the GML text, it attaches the spatial tags like feature types and coordinates to information to store in spatial database tables and non-spatial tags like attribute in non-spatial tables. In addition, the two tables have one-to-one relationship. That is, each row of spatial data table associates with one and only one row of the non-spatial table. The problem encountered here is that a GML document can involve multiple spatial and non-spatial attribute for individual feature; it can also store multiple features. The solution that we propose to solve this problem is using the JSON format. This means that for every feature all spatial data is stored in a JSON file and all non-spatial attribute similarly is stored in a separate JSON file. In this way, each feature has two JSON file, one of them stores non-spatial information and the other one stores spatial information, each of these files are stored in their own tables. Thus, the one-to-one relationship problem of spatial and non-spatial information is solved. These two tables and URL's table initially stored for GML documents have a one-to-many relationship. This means that each data record in the URLs table can be linked with multiple records of spatial and non-spatial tables, because as noted above, each GML file can contain several features. In addition to these three tables, there is a table to store the GML documents track in pilot version which it will be eliminated in the final version, because in final version GML documents located in real Web space. The mentioned table and URL's table have a one-to-one relationship.

3-2 Online phase

1-3-2 User interface design

The user interface provides interaction between user and system. In suggested system user interface consists of two search methods. In keyword-based method, users import their keywords in user interface and after completion of the search process user interface displays results to them.
In coordinate-based method, users can make their searches in geographic coordinates (longitude and latitude) or UTM projection coordinates (x and y) mode. In UTM mode, the user must import zone number in the user interface.

2-3-2 Ranking mechanism

Ranking mechanism operates in two different search methods as below:
In keyword-based method, the ranking is based on similarity of user target keyword and words exist in documents. In such a way that documents that have more similar words with target keyword earn higher rank in results.
In coordinate-based method, ranking is based on distance between target point and features in documents. In such a way that documents that have a feature with less distance to target point earn higher rank.

3 CONCLUSION

Our proposed system as a spatial search engine that provides the possibility of searching throughout the GML documents and thus it improves the efficiency of spatial search engines. Since GML documents include explicit spatial information along with non-spatial information, the main advantage of this system compared to other spatial search engines is an integrated approach to spatial and non-spatial data.
We plan to extend our work to cover other documents containing spatial data such as KML, GeoJSON and TopoJSON in order to be able to present augmented results. Also we plan to optimize the crawler to reach better search results on actual data on the Web.

4 REFERENCES

Bone, C., Ager, A., Bunzel, K., & Tierney, L. (2014). A geospatial search engine for discovering multi-format geospatial data across the web. International Journal of Digital Earth, (ahead-of-print), 1-16.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248-263.

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., & Vaid, S. (2004). The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In Geographic Information Science (pp. 125-139). Springer Berlin Heidelberg.

Li, W., Yang, C., & Yang, C. (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.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., ... & Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. International journal of geographical information science, 21(7), 717-745.

Walter, V., Luo, F., & Fritsch, D. (2013). Automatic Map Retrieval and Map Interpretation in the Internet. In Advances in Spatial Data Handling (pp. 209-221). Springer Berlin Heidelberg.

Wang, C., Xie, X., Wang, L., Lu, Y., & Ma, W. Y. (2005, November). Detecting geographic locations from web resources. In Proceedings of the 2005 workshop on Geographic information retrieval (pp. 17-24). ACM.

Wang, X., Zhang, Y., Chen, M., Lin, X., Yu, H., & Liu, Y. (2010, June). An evidence-based approach for toponym disambiguation. In Geoinformatics, 2010 18th International Conference on (pp. 1-7). IEEE.

Zhao, J., Jin, P., Zhang, Q., & Wen, R. (2014). Exploiting location information for web search. Computers in Human Behavior, 30, 378-388.