RESEARCH ON PROBLEM-BASED SPATIAL AND NON-SPATIAL INFORMATION SEARCH METHODS

Hongsheng Li [a, b, c, *], Jiping Liu [c], Yong Wang[c], Qingyuan Li [c]

[a] School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China- lihongsheng@163.com
[b] Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China- lihongsheng@163.com
[c] Chinese Academy of Surveying and Mapping, 16 Beitaiping Road, Beijing 100039, China-(liujp, wangyong, liqy)@casm.ac.cn

KEY WORDS: Data Search Methods, Spatial Information, Non-Spatial Information, Integration and Visualization, Text Extraction

ABSTRACT:

In the field of information processing, there exists awkward comparison between information flooding and thirst for information. One of the most publicized goals of modern information processing is to provide flexible access to information for anybody, anywhere, anytime. How to get needed data or information from vast information ocean, especially problem-based data search is becoming a focused research domain nowadays. Experts from Geography Information System community as well as Information Retrieval community are collaborating in order to fill and level up this great gulf. Driven by user requests and technology trends in the field of IT (Information Technology) and NSDI (National Spatial Data Infrastructure), this paper presents an ongoing research that intends to explore approaches to problem-based spatial and non-spatial information search method. The research methodologies and system flow chart of search prototype are well-described. A specific focus of this research is the intelligent parsing (formulizing) method of user's problems and subsequent information search method from diverse and heterogeneous data sources. Spatial and non-spatial information post processing methods are also discussed for the purpose of elegant search precision. Effective approach of integration and visualization of spatial and non-spatial information is mentioned. Experiments on extraction of geographic information from plain text are carried out to validate the feasibility of our approach based on text categorization and extraction theory. Conclusions and future works are introduced in the end of this paper.

1. MOTIVATION

As Nobel Prize winning economist Herbert Simon wrote nearly forties years ago, "The most scarce resources is no longer information itself, but the capability of processing information in the information era."(Herbert A. Simon, 1966) However the first thing is to get information that you want from vast information ocean. The information era creates new opportunities as well as challenges for information retrieval. The amount of information is growing rapidly, as well as new users inexperienced among which many users encounter the information starvation (Sergey B., Lawrence P., 1998). Search method or search engine can be the bridge of communication between information ocean and user request. However search engine nowadays can not meet people's request in the geographic domain for the complexity of spatial entities. As the geographic data comes into more valuable application, more and more people want to integrate the spatial data, attribute data, even temporal data in order to support spatial decision or to explore geographic pattern.

A new research proposal comes into consideration: problem-based spatial and non-spatial information search method. The original idea comes from our honoured user on the project approval meeting at the end of 2004. "If user has some spatial and non-spatial problem, he or she may surf the net or retrieve in local databases to find answers and even some solutions." Some questions are easy and there are definite answers, while others need to be deeply analysed and more solutions need to be explored. The search system prototype should comprise simple questions concerning non-spatial information like "Who is Bill Gates?", simple questions concerning spatial information like "How far from Beijing to Amsterdam ", and complex questions concerning both spatial and non-spatial information like "How can I leave Beijing to Amsterdam?". The last question needs to integrate spatial layer like railway, airline as well as descriptive non-spatial data like transportation fare, social custom etc.

2. BACKGROUND AND RELATED RESEARCH

Search engines are main tools for people to retrieve information on the web nowadays. The industry has grown up around the widely published opinion that: " Knowledge workers spend 35% of their productive time searching for information online, while 40% of the corporate users report they cannot find the information they need to do their jobs on
the internet."(R. Baeza Yates, B. Ribeiro Neto, 1999). Most search engines are based on key words query with limited scope covered and low successes rate. It can support full text search but can not handle semi-structured, multi-media and specifically spatial information. Intelligent and semantic-based search engines are still in leading strings (Hugo Liu, et al.. 2002; M. Sintek, S. Decker, 2002; Jukka Perki¨o, et al.., 2004).

As NSDI calendared in more country and geographic information request grows, GIS users want to have access to more-powerful tools to find geographic information from widely disparate sources. An American patent (No. 6,772,174) titled with "Data administration method" shows that search

* Corresponding author.

engine is advancing in the field of spatial data search. Spatial data layer and attribute data relating to certain spatial entity can be retrieved based on geographic location. This approach can be categorized into Geographic Information Retrieval (GIR) (Mário J. Silva, et al..2004; Jones, C., et al..2002). To make Canada's geospatial databases available on the internet, Canada has published a white paper to promote developers to be aware to make their spatial data and even applications be searchable (GeoConnections Secretariat, 2004.). As Peter L. Croswell forecasts, "Major improvements in geographically based queries against unstructured data types (use of data catalogue systems, full text search, XML databases, emerging semantic Web standards) for geographic data exploration and retrieval will be realized" in the coming ages. Innovations in search engines and web intelligence will support efficient and precise data searches. "Current web architecture and search tools will go beyond queries on text and keywords—meaning and context can be examined, with more sophisticated search algorithms and greater use of XML and the Resource Description Framework (RDF)."(Peter L. Croswell, 2005.).

3. OUR RESEARCH FRAMEWORK AND METHODOLOGY

3.1 Research objective

Unlike GIR which is concerned with retrieving documents that are related to some location (Marc V., et al. 2004), unlike NSDI which make pure map and related data search conveniently, our research objective is construct a seamless and integrated prototype system. The system can locate geographic terms based on user problem or request, and search related diverse spatial information and non-spatial information. What's more, it can present these two kinds of heterogeneous information integrative.

3.2 Research Methodologies

Problem-based spatial and non-spatial information search will be the bridge between vast information and user request and can be recognized as a kind of information mining process guided by users' problem. Our methodology can be concluded as three parts: divide and conquer, integration, dynamic visualization. Spatial information and non-spatial information have different characteristics. For non-spatial information especially full text, many search engine like Google™ can be the candidate. However few search engines nowadays bring spatial information into consideration. Spatial information search method is most difficult and should be given more priority. The strategy "divide and conquer" is to divide information into spatial information and non-spatial information and conquer each of them separately. This approach can make both aspects easily processed. However when the answers are fed back to users both aspects should be integrated in order to get relevant information well understood and to support decision. The answers need to be dynamically visualized so as to be friendly to users and adaptive to the refining search process.

3.3 System flow chart

Deeply analysed the research objectives, our primary design provides a conceptual system flow chart. Detailed description for Figure 1 is given as follows.

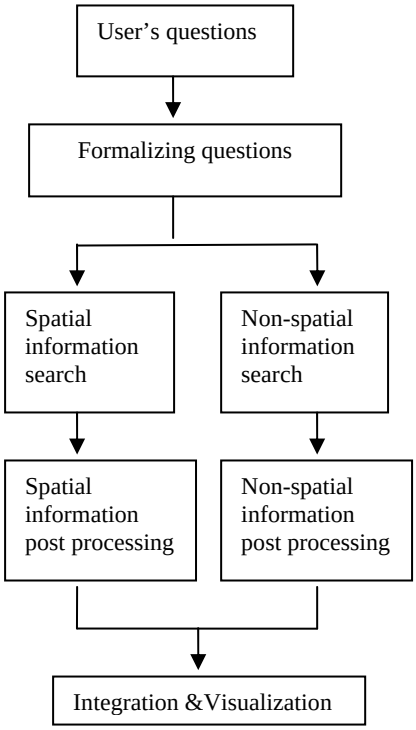

Figure 1 system flow chart

3.3.1 Users submitting their problems

Based on their specified problems or questions, users submitted their request for information. The prototype system should be robust enough to endure vicious input. Users can also submit text file that can explain more precise situations.

3.3.2 Formalizing the questions

The aim of this step is to make the computer understand semantic meaning contained in the question. Following sub-steps are needed.

1. Categorization of user request: Questions and problems should be categorized into several types for simplicity. Respective answers and solutions are matched by means of regular expressions or text categorization.

2. Feature extraction: Text categorization is the precondition of text extraction. This sub-step is to extract some descriptive information of user request, especially keywords. To extract or explore basic features of user's input is truly a text mining process. Text extraction as well as text categorization is researched in two different approaches: One is regular based text categorization and extraction, the other is statistical based text categorization and extraction (Dejun X., Maosong S. 2004). And nearly-mature regular based approaches will be given more focuses. The key is to construct adaptive text categorization and extraction rules for geographic information. A lot of algorithms such as intelligent Chinese character split algorithm, text categorization regular training algorithm, text categorization algorithm and text extraction algorithm are needed.

3. Spatialization of geographic information: If there exists geographic or spatial information, they mainly are place names or qualified geographic location such as 20 km north to Beijing, China. If not, the spatial information search may be not necessary in this case. The spatialization of geographic information is then needed. Spatialization will map geographic information into spatial extent or coordinates (in latitude/longitude or other unified coordinates system). However resolving the definition of certain geographical terms is still an active research area. Terms that describe so-called "ill-defined regions" such as the "Midwest" in US or the "Midlands" in the UK are such an example. After delineation, the regions can be stored as geographic features in a database or an ontology (Avi Arampatzis, et al.. 2004; Markowetz, A., et al..2003). On the other hand, place names that varying with time or conflicts between different place name standards are also concerns.

4. Identify the user request: The results of this step are adapted keywords couples like <key, value> that following specific self-defined XML format. Types and content of user's problem, spatial location (like MBR-Minimum Bounding Rectangle), and possible answer type are some basic features for describing the user request.

3.3.3 Spatial information search

After the above two steps, prototype system has already know what to search. However how to search is more important. It is divided into two aspects, spatial information search and non-spatial information search. Spatial information search is the most technically intensive step. It mainly contains following terms.

5. Spatial information search in homogeneous system;
6. Spatial information search in heterogeneous systems;
7. Spatial information search in local spatial database;
8. Web spatial information search;
9. Matching and integration of local and web spatial datasets

GIS data portal and nodes in the geographic data infrastructure can be accessed by means of metadata or directory. Many geographic information service sites can provide spatial data services from raw data, static maps, vector maps to bundle of datasets; other sites may further provide spatial operation services from simple query to complex buffer, route analysis which conform to OGC web service specification. These sites can be utilized to extract relevant spatial information by means of OGC web service like WMS, WFS or WCS etc. (Gong J. Y. etc.2004.) More related topics can be found in the field of geographic interoperation. And the service quality needs to be evaluated according to user's request. Our primary design intends to be open framework on the base of open standards such as GML, OGC, and web services technology. One alternatives is to implements a simple adapter or wrapper on each heterogeneous application systems that will provide geographic services like WMS, WFS etc. Another alternative is to research formalized representation theory and method of software system and as well as spatial and non-spatial datasets in the software system by means of metadata or other methods.

3.3.4 Non-spatial information search

Local databases should be queried firstly for its convenience. If the system can not find enough information, it will turn to web search. The primary aim of web search is to integrate QA (Question and Answer) system and meta-search engine. If user just wants to know simple questions about certain fact, the prototype system will submit query to some QA systems. If complex questions are asked, then the prototype system will submit query to search engines after reconstructing some keywords based on specified grammar of certain search engine. The Meta search engine is based on successful search engine like Google™ , MSN™ search etc. for English text and Baidu™, Zhongsou™ etc. for Chinese one. Different query condition should be assigned different weight based on different characteristics of these QA systems and search engines when integrated.

3.3.5 Spatial information post processing

This step is mainly based on user's request. It contains spatial analysis method from simple length or area calculation to complex overlay analysis, buffer and optimised route query. Related algorithms can be found in the field of geographic information system.

3.3.6 Non-spatial information post processing

This step mainly concerns non-spatial information organization and web page link analysis. To make sure the possible answer easily acquired by users, web page ordering, text categorization, site clustering and content clustering, display style optimisation should be thoroughly considered. The total process should be carried out by means of feedback mechanism and progressive refinement. Related theory and method should turn to information processing field.

3.3.7 Dynamic integration and visualization

Effective approach of integration and visualization of spatial and non-spatial information should be explored. This is where the complementarity of the spatial approach and non-spatial approach so acutely emerges. And at the same time the comparison of both can be helpful to authentic application. This step is a kind of dynamic process and the prototype system should be aware to changes of both spatial information search results and non-spatial ones.

4. ACHIEVED RESULTS

Formalizing questions from a geographic point of view suggests that categorisation and extraction of spatial-related text. Our approach is to explore attribute, spatial and temporal information from natural text. Lots of experiments have been done to test its feasibility. Text categorization and extraction rules for geographic information are constructed. Text-related spatial location can be easily located on the map for intuitionistic analysis and decision.

4.1 Text categorization and extraction

In reality, a substantial portion of the available information is stored in text databases. And data in most text databases are semistructured. In our experiment, only plain text is well researched. Figure 2 shows the flow chart of text categorization and extraction. Firstly, the raw text data will be converted into text database. According user requests, key words for text

categorization are well-chosen. And after sample text training, categorization rules are generated. According these rules different texts can be categorized for extraction of specified information (location and attribute) utilizing latitude, longitude, place names, azimuth words(e.g. “southwest”, “centre” etc.), .placement words(e.g. “located”, “from”, “to” etc.), distance etc.

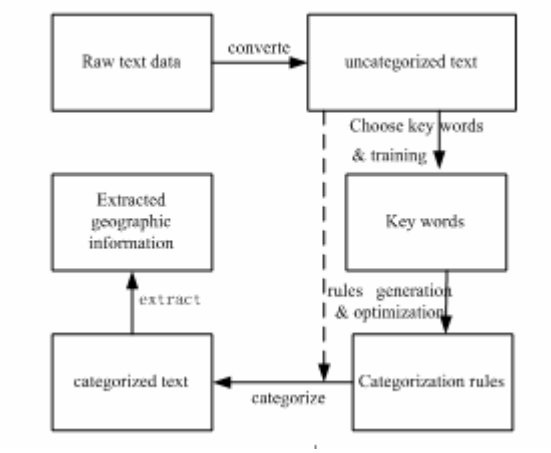

Figure 2 Flow chart of text categorization and extraction

4.2 Experiments

Follows are some snapshots of our experiment prototype. Figure 3 illustrates the edition process of key words for text categorization. Figure 4 shows generating process of categorization rules.

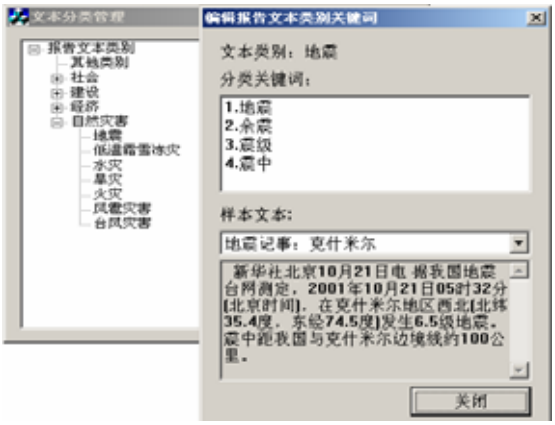

Figure 3 key words edition

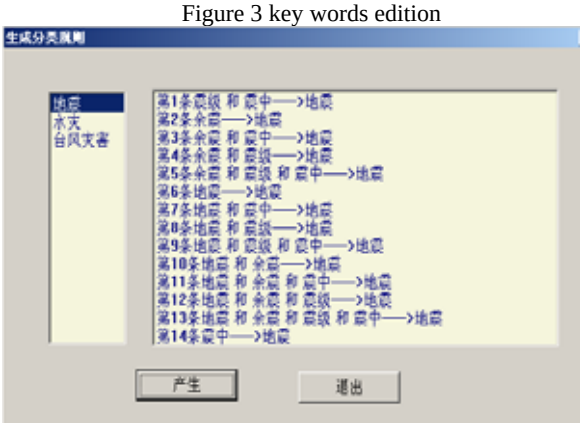

Figure 4 generating categorization rules

Figure 5 displays the association of extracted place names (earthquake centre) and bounding box of the specified geographic entity. The text says that an earthquake happened and its earthquake centre was located at TaoYuan Town, Yong Sheng County, Yun Nan Province. According the spatialization process described above, the bounding box of earthquake centre is displayed on the map. Designer will benefit from map-like representation to support exploration of geographic patterns and effective solution design. Spatial information search in this experiment only considered homogeneous and local spatial database. Spatial information search from heterogeneous spatial database and from web servers is more demanding and necessary in interoperable environments. Meantime non-spatial information search and its refinement are also our next research direction.

Figure 5 associations of geographic terms and entity

5. CONCLUSIONS AND FUTURE WORKS

5.1 Conclusions

A first innovative aspect of this multi-discipline research is the introduction of search method into geographic domain. It extends the Geographic Information Retrieval (the search results are documents) and map data search (the search results should be map related data). The search outcome integrates both documents and map data. Only in this way can users compare and verify two kinds of information from different approaches.

A second innovative aspect is that application pattern of spatial and non-spatial data based on user’s problems can be widely used in different application field whenever lack of information, such as risk response system, disaster review and analysis,

personal travelling system etc. It can be easily extended to support topic-based search, event-based search and so on. The research is expected to play a major role in filling up the current gap between user expectation and requirements for information, and the vast information ocean that is growing at high speed and not be utilized by people.

5.2 Future works

The research comes from truly user request and is in accordance with technological trend. Four sub-tasks has been already identified and are on their way, i.e. formalized description of spatial problems, homogeneous and heterogeneous spatial dataset access method, local and web mining method and dynamic integration and visualization of search results. This prototype system will constitute an initial answer for supporting the process of search and integration of existing diverse information sources in heterogeneous environments. It plans to support acquisition of information about user specified requirements. The research will build an integrated representation of related information from the point of view of geographic information system and information retrieval. The final objective is to construct an intelligent, integrated and full-functional information search system prototype which is the bridge of communication between vast information ocean and users' request.

ACKNOWLEDGEMENT

This research is sponsored by Key Laboratory of Geo-informatics of the State Bureau of Surveying and Mapping.

REFERENCES

Avi Arampatzis, et al. 2004. Web-Based Delineation of Imprecise Regions. In Proceedings of Workshop on Geographic Information Retrieval, SIGIR, Sheffield, UK.

Dejun X., Maosong S. 2004. Eliminating High-Degree Biased Character Bigrams for Dimensionality Reduction in Chinese Text Categorization. Proceedings of ECIR-04, *26th European Conference on Information Retrieval Research*, Springer Verlag, Heidelberg, DE. pp. 197-208.

GeoConnections Secretariat，2004., A Developers' Guide to the CGDI: Developing and publishing geographic information, data and associated services,

http://www.geoconnections.org/CGDI_Technical_Manual_0204_e.pdf. (Accessed 28 Jan. 2005)

Gong J. Y. etc.2004. *Modern Geographic Information Technologies*. Science Press. Beijing, China.

Herbert A. Simon, 1966. The Impact of the New Information-Processing Technology, *Economy*.

Hugo Liu, et al. 2002. GOOSE: A Goal-Oriented Search Engine with Common-sense. In De Bra, Brusilovsky, Conejo (Eds.): *Adaptive Hypermedia and Adaptive Web-Based Systems*, Second International Conference, AH 2002, Malaga, Spain, May 29-31, 2002, Proceedings. Springer, pp. 253-263.

Jones C., et al., 2002. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project [C]. In proceedings: 25th ACM Conference of the Special Interest Group in Information Retrieval, pp387-388.

Jukka Perki¨o, et al., 2004. Exploring Independent Trends in a Topic-Based Search Engine. HIIT Technical Reports. pp. 1-7

Markowetz, A., et al. 2003. Exploiting the Internet as a Geospatial Database. ISPRS WG IV/5 Workshop on Next Generation Geospatial Information. Cambridge, UK.

Mário J. Silva, et al., 2004. Adding Geographic Scopes to Web Resources [C], In Workshop on Geographic Information Retrieval, SIGIR '04, Shefeld, UK.

Marc V., et al. 2004. Distributed Ranking Methods for Geographic Information Retrieval，20th European Workshop on Computational Geometry March 25-26, Seville (Spain)

M. Sintek, S. Decker. 2002. Triple: A Query, Inference and Transformation Language for the Semantic Web [C]. ISWC'2002, Sardinia, pp. 364-378

Peter L. Croswell, 2005. Technology Trends and Opportunities for the Future, http://www.geoplace.com. (Accessed 28 Jan. 2005)

R. Baeza-Yates., B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

Sergey B., Lawrence P., 1998. *The Anatomy of a Large-Scale Hyper textual Web Search Engine*. Computer Networks and ISDN Systems, vol30:pp107--117.