ON THE FRAMEWORK AND KEY TECHNOLOGIES OF MODERN GIR SYSTEMS

Xing Lin and Yifang Ban

Division of Geoinformatics, Royal Institute of Technology (KTH)
Drottning Kristinas väg 30, 100 44 Stockholm, Sweden
mailto:{xingl, yifang}@infra.kth.se

Commission WgS – PS: WG II/2

KEY WORDS: Information Retrieval, Spatial Information Science, GIS, GIR, Framework, Semantics

ABSTRACT:

This paper intends to present a new conceptual framework for a modern GIR system as well as its related key technologies. In this research paper, existing GIR systems were surveyed and compared to find out the challenges we are facing in GIR and corresponding research needs in the future. New ideas on how to solve these existing problems are proposed in this paper, including a new system framework and promising key technologies. In addition, recent progress and hot technologies in geographical information science and information science fields will be presented, which will certainly be useful to build a better GIR system. It is believed that the research and discussion presented in this paper could, to some degree, facilitate the research and implementation in the field of GIR.

1. INTRODUCTION

Geographic information (GI) is one of the most important and the most common types of information's in human's society. Recenct efforts have been made either by expanding the traditional IR to support a spatial query (Baeza-Yates and Ribeiro-Neto, 1999), or building a GIR in a brand new architecture from the ground such as the SPIRIT project (Jones et al., 2002; Jones et al., 2004; Purves et al. 2007). To some degree, these existing GIR systems could solve users information research need with a spatial filter, especially when the users are looking for information on something within a relatively big extent, for example, hotels in Stockholm. When you submit the "Hotel" and "Stockholm" as keywords to popular web search engines, one could also find most relevant information of one's interest. According to Jones and Purves (2008), existing approaches, however, have many shortcomings and need to be improved to build better GIR systems in the following aspects: a) detecting the geographical information within users queries and text documents; b) disambiguating the place names to find the intended one; c) interpreting the geometric location of vague place names; d) spatially and thematically indexing the text documents within a GIR system; e) information retrieval model to pickup the relevant documents out of the library and ranking the degree of relevance according to their spatial and non-spatial properties; f) effective user interface; g) approaches to evaluate the success of a GIR system. Some of them might require new techniques to be applied, while others might rely on a better system architecture.

Therefore, improvement still needs to be made to existing solutions. The objective of this paper is to present a discussion on a better solution to these issues by improving existing solutions or making a new one. In respond to those problems, solutions could be concluded as the following five aspects: (1) A proper representation model and extractor of geographical information for text documents based on ontology; (2) An innovative information retrieval model and relevance ranking algorithm; (3) A combined indexing mechanism for both geographical and thematic content; (4) A new GUI integrating digital maps and text contents; (5) A new system architecture that enable the whole system to self-learning and evolve.

Hence, this paper is divided into four sections. The first section (Section I) will give a general introduction to the purpose and of this paper. Then, related key techniques will be presented and discussed in the second section (Section II). In Section III a new architecture is proposed for modern GIR systems, which will incorporate all those related key techniques introduced in the previous section. The last section (Section IV) presents conclusions and future work.

2. KEY TECHNOLOGIES AND RESEARCH NEEDS

In the following part of this section, some of the most important key technologies that modern GIR systems need will be covered, based on previous research and the authors' own opinions.

2.1 A proper representation model and extractor of geographical information for text documents

In the past 10 years, the digital gazetteer has been playing a very important role in the research and development of GIR. Based on the domain ontology of GI, geographical thesaurus and digital gazetteers, recent GIR research has adopted an integrated approach to represent, detect and estimate the geographical information from documents of natural language (Alani et al., 2001; Jones, 2003; Ø Vestavik, 2004; Jones, 2004; Fu et al., 2005; Mata 2007). Based on the contextual properties of named places, co-occurrence model of places within the same document has been invented to help disambiguating the places, which might refer to totally different places within different context (Overell and Rüger, 2008). A good example of this problem is the named place of "London" as a city in Ontario of Canada or the capital of UK.

Although the use of digital gazetteers brings a big improvement to the performance of modern GIR systems, conventional digital gazetteers still have some problems.

(1) One the problem is the vagueness of geographical information within the named places and spatial qualifier of natural language. The vagueness might include different geographical boundaries, different feature types, data content

changing over time, vagueness caused by abbreviation and so on. Such vagueness or uncertainty comes from the vagueness of human language when communicating geographical concepts between each other (Liu et al., 2007a). There are some named places that are widely used but never have a fix geometric boundary. Although there is no fence surrounding the campus of KTH (see Fig. 1), people who are familiar with KTH always have a rough but fixed boundary of KTH in their common sense. Those people won't face any problem when talking about events or buildings inside or outside KTH during their conversation. But this might cause problems for computers to process the queries like "restaurants inside KTH", because a fixed and numerical boundary of KTH is needed for computation.

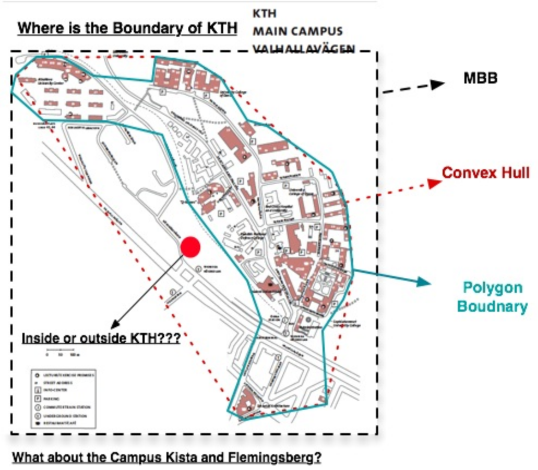


Figure 1. Where is the boundary of KTH?

(2) Besides, there are a large number of temporary or implicit places existed in the text documents, for example, "the building is located just between the university and the stadium". Traditional gazetteer doesn't cover such kind of places, and thus can't provide geo-references for them.

(3) The interpretation of these vague spatial qualifiers might depend on the scale and shape of reference objects. For example, the "near" qualifier is the most commonly used spatial predicate by the GIR users since people always care more about the things and events happened in their locality. But in reality, it is very hard to assign a proper distance from reference object to represent the "near" qualifier. Besides, there are some other spatial qualifiers are more qualitative rather than quantitative, as illustrated in Fig. 2 and Fig. 3.

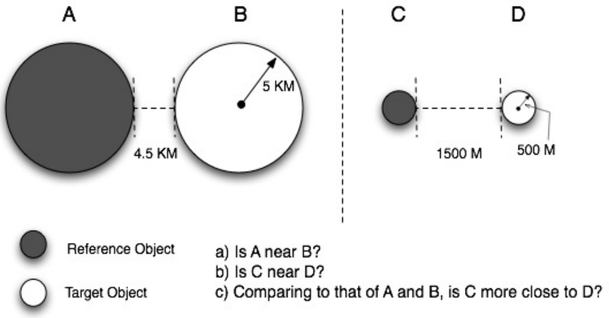


Figure 2. Different interpretation of spatial qualifier "near" in different scale and reference object

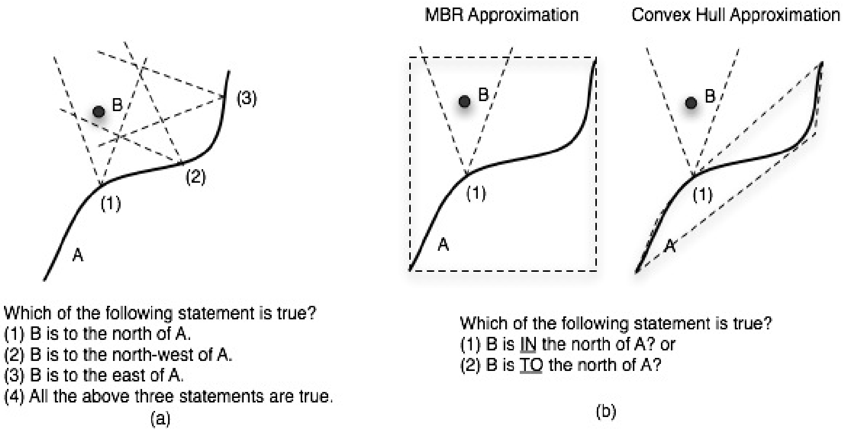

B
(3)
(2)
(1)
A
Which of the following statement is true?
(1) B is to the north of A.
(2) B is to the north-west of A.
(3) B is to the east of A.
(4) All the above three statements are true.
(a)
MBR Approximation
Convex Hull Approximation
B
(1)
A
B
(1)
A
Which of the following statement is true?
(1) B is IN the north of A? or
(2) B is TO the north of A?
(b)

Figure 3. Shape and geographical approximation will affect the interpretation of cardinal direction

From the above discussion, it could be concluded that modern GIR need to deal with quite a lot of query and geographical information in a more qualitative way, rather than the quantitative way from the specialized GIS. One of the promising representation models is the Generalized Place Name (GPN) model proposed by Liu et al. (2007a). Comparing to the simple named places model used in digital gazetteers, the GPN could directly or implicitly geo-reference any text content that describes a certain location on earth surface. The most important advantages of GPN versus traditional digital gazetteers are: (1) Not only named places, but also all the other geo-references could be collected and represented in the GPN library; (2) Not only the geometric boundary, but also the qualitative spatial qualifiers are remained in the GPN library to enable spatial reasoning. (3) By introducing the scale and timestamp, the geographic footprints in GPN are also more advanced than the digital gazetteers. Contextual information is included to solve the ambiguity among places share the same names.

But obviously, the GPN will bring more geo-parsing and storage overhead when introducing it to modern GIR systems. Typically, most GIR systems are built as an extension to existing IR systems. How could the GPN library be integrated into and used by the traditional IR systems still requires further research and development. The structural complexity of GPN will also certainly require more computational resource. In addition, in GPN library, there will be more than one geo-reference for a single document. These geo-references, no matter inside the same document or between documents, are connected to each other via spatial relationship and hence form a huge network. A quick indexing mechanism and search algorithm is especially needed to find a GPN through out the network.

The authors believe that a possible way to build a GIR system using GPN library is that the thematic and geographical retrieval processes are executed respectively. Then the result is combined and ranked according to both the similarity score in thematic dimension and geographical dimension. Concerning that fact that there will be more than one geo-footprint in a single document, the co-occurrence model of these geo-footprints could be studied to estimate the importance of each places towards the document. We need to know not only the geo-footprints of documents, but also the importance of each geo-footprints. In other words, it needs to be found out that which place is this document mainly talking about. Together with the geographical similarity, this importance score could then be used to rank the geographical relevance between the query and document. More about this new information retrieval model and ranking algorithm will be introduced in the coming section.

2.2 A innovative information retrieval model and relevance ranking algorithm

In an eligible GIR system, all related documents stored in the system should be returned, while the most important and concerned ones should be put in the front rows. To achieve such a goal, the information retrieval model, which estimates the similarity between document and query, is the vital problem. This problem has been well resolved in the thematic perspective, which leads to the great success of Google and Yahoo. But its geographical counterpart is far from satisfaction. In existing GIR systems, the single or overall geographical footprint is adopted. In single geographical footprint model, the chosen geographical reference of those appeared in the text document will be adopted as the geographical boundary for this document. The one being chosen could be the most important (frequent) one or a random one. In the overall geographical footprint model, the geographical boundary of a text document is produced by aggregating all those geographical references, which exist within the same document. Among various kinds of single/overall geographical footprint model, the most important approaches are the centroid, maximum bounding box (MBB), convex hull and generalized polygon boundary.

During the previous efforts of GIR in the aspects of geographical similarity measurement, most approaches are based a single geometric spatial properties, either the intersection of the query's and the document's geographical boundaries or the Hausdoff distance. Based on the geographical footprints encoded in MBB, the first approach is quite easy to understand and implement. The Hausdoff distance approach not only considers the area of intersection, but also takes into account the similarity of shapes between query's region and document's region, as well as the distance between them (Frontiera et al., 2008).

Besides the simple geometric measurement based approach, there is also other more complex but with better performance algorithm for spatial similarity calculation. The probability ranking approach is one of the most famous. With the help of logistic regression, an equation could be established to map the factors, which might affect users' decision of relevant document, to the degree of relevance. Based on manual interpretation, a training dataset, queries and preferred answers could be prepared to help determining the parameters in the logistic regression formula (see Equation. 1).

$$P(R \mid X) = \frac{1}{1 + e^{logO(R|X)}} \quad and \quad logO(R \mid X) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad (1)$$

X_i are the factors that might affect the final probability of being relevant document, and the β_i are the parameters for corresponding factors. Useful factors could be the ratio of overlapping area against the area of query, the ratio of overlapping area against the area of document, and so on. The main advantage of this probability ranking approach could provide an optimal or a near-optimal retrieval performance, and the ability to use statistical methods with meaningful indicators for both the design and evaluation purpose of a GIR system. According to Fronteria's result, the probabilistic approach could better estimate the spatial relevance than the two single geometric approaches mentioned before (Frontiera et al., 2008). Although the single or overall (aggregated) geographical footprint model is simple, straight, and quick, this might cause the problems of overestimation and underestimation of geographical scopes for the text documents.

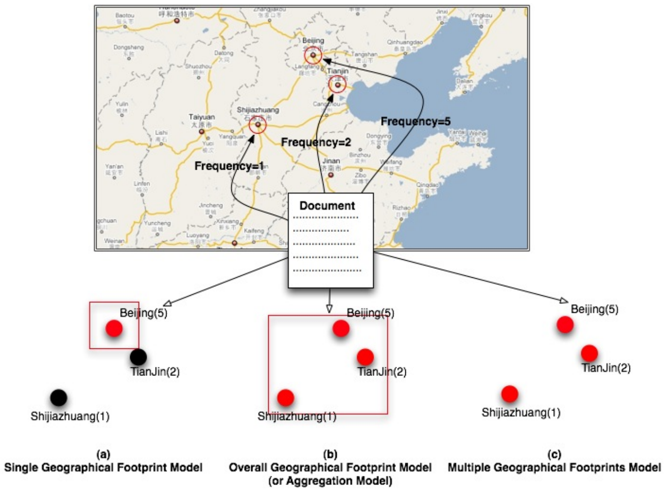


Figure 5. Geographical footprints models: (a) single model; (b) overall/aggregated model; (c) multiple model

From the author's point of view, the multiple geographical footprints model should be applied to represent the geographical information contained in a single document. The inter-impact of these spatially related geographical entities within a document is worthy of being studied to help the process of estimation. In addition to the spatial similarity between query and documents, the degree of importance of each place to its hosting document should also be put under consideration. Since there are several geographical references in a single document, not all of them are equivalently described within the context of hosting document. The following aspects are worthy of consideration: (1) The most obvious indictor to estimate the degree of importance is the frequency of occurrence; (2) Besides, the occurrence of one named place could serve as an evidence for the occurrence of another named place, which is spatially related to the previous one. Such impact varies with the strength of different spatial relationship between geographical references within the same document; (3) What's more, the importance of a certain named place will also influence the inter-impact it casts on the other places. If place Solna is quite important for document A and in the same document there is another named place "Stockholm" which contains Solna, it is safe to say that the document A should also have good probability to be talking about something in the city of "Stockholm". These three principles provide a basis for evaluating geographical importance for each single unique geographical reference within a document. Hence, in the multiple geographical footprint model, the geographical adjacency and geographical importance will finally be merged to generate the final score of geographical relevance.

Another important problem in IR model is the combination of the two similarities in the thematic dimension and geographical dimension. The most popular way to combine these two similarities is a weighted linear combination of both similarities demonstrated in the following equation (Equation. 2).

$$Re\,l(q,d) = \omega_T * Re\,l_T(q,d) + \omega_G * Re\,l_G(q,d) \qquad (2)$$

But static weights of ω_T and ω_G might cause problem because the weights might vary among different scenarios of queries.
Generally speaking, it is nearly impossible to find two static weights that could be suitable for any case of queries. In fact, the weights have great concern with the specialty of user queries, both in the geographical aspect and the thematic aspect. What's more, different people might have slightly different point of view on which component is more important than the other. Some research has been devoted to study the specificity of queries (Yu and Cai, 2007), which will dynamically affect the weights for each query. Basically, this is a promising approach that could solve the problem of similarity combination to some degree. Beyond the weighted linear combination, there are also some other ways to produce a final score of relevance based on the two in thematic and geographical scopes. Based on the fact that these two scopes are independent from each other, the geometric average of scores in these two dimensions could be calculated as the final value of relevance ranking.

$$Re\,l(q,d) = \sqrt{Re\,l_T^2(q,d) + Re\,l_G^2(q,d)} \qquad (3)$$

From the author's point of view, the combination of similarity scores needs to consider the following four factors.

- $Re\,l_T(q,d)$ - Similarity score in thematic scope;
- $Re\,l_G(q,d)$ - Similarity score in geographical scope;
- S_T - Specialty factor of query term in thematic scope;
- S_G - Specialty factor of query term in geographical scope;

Similar to Equation 1, the same logistic regression approach could be applied here to figure out the function of combination. During the logistic regression process, two feature variables will be:

- $X_1 = Re\,l_T(q,d) * S_T$ - Thematic component
- $X_2 = Re\,l_G(q,d) * S_G$ - Geographical component

The S_T value could be calculated from the hierarchical position of corresponding query term in a chosen thesaurus. At the meanwhile, the S_G value could be derived from level of places to query located in the geographical gazetteer. A widely collected training dataset could then be used to work out this logistic regression. It is believed that this approach could generate a more proper combined score of similarity out of the two individuals in respective scopes. More effort of research is worthy to be put in this aspect.

2.3 A combined indexing mechanism for both geographical and thematic content

Indexing technologies enables the fast retrieval of related document from the document library. The inverted file structure is the most important indexing tool based on key words for modern web search engines to find chosen document in a short time. But for the GIR systems, the documents need to be indexed based on both the thematic and geographical features.

The inverted file structure (IFS, or referred as posting files) is the dominant indexing mechanism, which is widely applied in modern web search engines. Inverted file structure enables quick full text search based on one or more key words as queries terms (Berry and Browne, 2005). Among the spatial indexing technologies, there are also some quite successful indexes, which have also been widely applied in GIS and Spatial database systems (SDB). Popular spatial indexing technologies are grid file, space-filling curve (e.g. z-order, Peano Curve, Hilbert curve), quad-tree, octree, kd-tree and R-tree family (e.g. R-tree, R+-tree, R*-tree), as listed in the following two figures. Among these spatial indexing technologies, the R-tree family is the most important one.

Concerning the approaches to combine thematic and spatial indexing technologies in modern GIR systems, there could be four different styles according to Lin et al. (Lin et al., 2007) for single geographical footprint model. They are: (1) Pure Keyword Index, PKI; (2) Keyword-Spatial Dual Index, KSDI; (3) Spatial-Keyword Hybrid Index; and (4) Keyword-Spatial Hybrid Index, KSHI. To find a proper combined indexing technique for modern GIR systems, the following aspects need to be taken into account: efficiency, storage overhead and operability. A trade-off should be made among these three aspects. Since most modern GIR systems using single/overall geographical footprint model are built as an extension to current traditional IR system, it is found that the KSHI (Keyword-Spatial Hybrid Index) is most suitable, which has a acceptable efficiency but least change to current index structure of traditional IR systems (Lin et al., 2007).

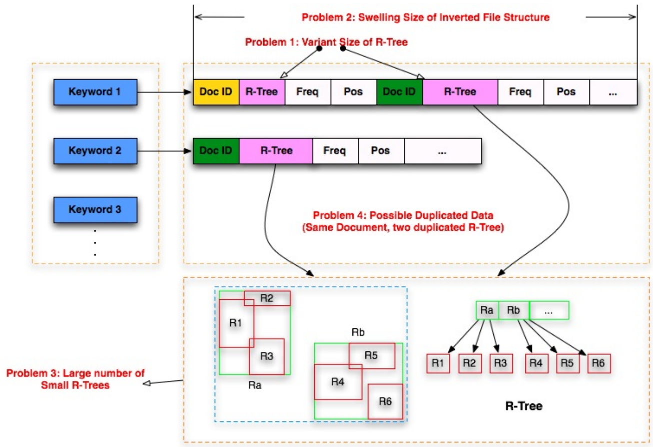

Problem 2: Swelling Size of Inverted File Structure
Problem 1: Variant Size of R-Tree
Keyword 1
Keyword 2
Keyword 3
Doc ID
R-Tree
Freq
Pos
Doc ID
R-Tree
Freq
Pos
...
Doc ID
R-Tree
Freq
Pos
...
Problem 4: Possible Duplicated Data
(Same Document, two duplicated R-Tree)
Problem 3: Large number of Small R-Trees
R1
R2
R3
Ra
Rb
R4
R5
R6
Ra
Rb
...
R1
R2
R3
R4
R5
R6
R-Tree

Figure 6. Potential problems while applying KSHI to GIR system using multiple geographical footprint model

But it is relatively more difficult to find a proper indexing mechanism for the multiple geographical footprint model, because there will be more than one geographical reference contained in a single document. The hybrid index of SFC and inverted files won't be applicable under this circumstance, since in this index structure there is only one slot allowed in the inverted file structure for the storage of geographical information within the same document. Although it is certainly allowed put one or more geographical references together with this single slot, then it might cause four problems: (1) variant length of geographical section; (2) swelling inverted file structure; (3) a large number of spatial indexes; (4) possible duplicated data (see Fig. 6). Hence the KSHI approach is not appropriate for GIR systems using multiple geographical footprint model, although it has been proved to be simple and efficient in single/overall geographical footprint model.

From the author's opinion, the more applicable index for modern GIR system of multiple geographical footprint model is the KSDI (Keyword-Spatial Double Index) because of the following four reasons.

(1) High indexing efficiency for information retrieval – In KSDI, an inverted file index and a spatial index are build respectively upon the whole document library. Although the dual index will bring extra storage overhead, it still has been proved to be able to notably improve the system performance of information retrieval (Lin et al., 2007).
(2) Least change to existing traditional IR system – While applying the KSDI, the indexing and information retrieval are carried out respectively in inverted files and spatial data storage. So it is obvious that KSDI will have nearly no change to existing traditional IR system. The only connection point between the two sub system is the operation of merging two preliminary result sets in thematic and spatial dimension. This process could be illustrated as the following figure (Fig. 7).
(3) No duplicated data storage – The geographical information within each document will only be stored, processed and indexed once in the KSDI index. No duplicated data storage will happen in such circumstance.
(4) Benefit from modern spatial database technologies – With the development of spatial database technologies, the SDBMS (Spatial database management system) has been the key components of modern GIS for spatial database management. Besides, with the help of modern SDBMS, you could easily add a simple digital map to your GIR system. In a word, the SDBMS has great potential to take over the task of spatial data management in modern GIR system. The modern GIR system will also benefit greatly from the appliance of modern spatial database technologies.

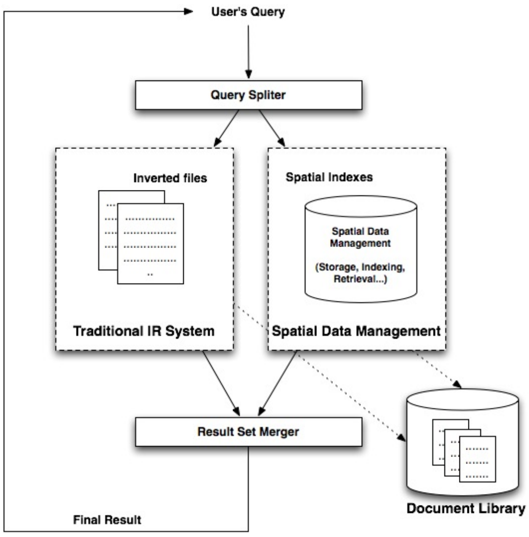


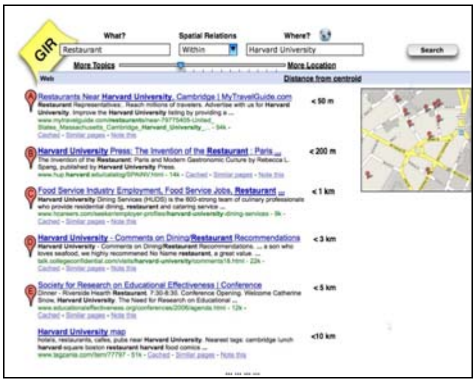
Figure 7. Query processing in GIR system using KSDI index

2.4 A new GUI integrating digital maps and text contents

The interface of an information system plays an important role to ensure the service quality and user experience by guiding the user to use the system in a proper way or make a better decision. A good user interface is also required for the success of a modern GIR system. In modern GIR systems, the user interface could be of great help in the following two aspects.

(1) <u>A good approach of rich representation for retrieved documents</u>

According to the author's opinion, a proper GUI for modern GIR systems might look like the following figure (Fig. 8). In this digital map powered user interface, brief text citation as well a rough location on digital map will be prepared for presentation for each retrieved document. The documents are sorted in a descending order by their score of relevance to the user's query. Estimated geometric measurement should also be presented according to different spatial relations used together with the named places.



Restaurant | Within | Harvard University

Restaurants Near Harvard University, Cambridge | MyTravelGuide.com — < 50 m
Restaurant Representatives. Reach millions of travelers. Advertise with us for Harvard University. Improve the Harvard University listing by providing a ...
www.mytravelguide.com/restaurants/near-79775405-United_States_Massachusetts_Cambridge_Harvard_University_... - 94k -
Cached - Similar pages - Note this

Harvard University Press: The Invention of the Restaurant : Paris ... — < 200 m
The Invention of the Restaurant: Paris and Modern Gastronomic Culture by Rebecca L. Spang, published by Harvard University Press.
www.hup.harvard.edu/catalog/SPAINV.html - 14k - Cached - Similar pages - Note this

Food Service Industry Employment, Food Service Jobs, Restaurant ... — < 1 km
Harvard University Dining Services (HUDS) is the 600-strong team of culinary professionals who provide residential dining, restaurant and catering service ...
www.hcareers.com/seeker/employer-profiles/harvard-university-dining-services - 9k -
Cached - Similar pages - Note this

Harvard University - Comments on Dining/Restaurant Recommendations — < 3 km
Harvard University - Comments on Dining/Restaurant Recommendations. ... a son who loves seafood, we highly recommened No Name restaurant, a great value. ...
talk.collegeconfidential.com/visits/harvard-university/comments18.html - 22k -
Cached - Similar pages - Note this

Society for Research on Educational Effectiveness | Conference — < 5 km
Dinner - Riverside Hearth Restaurant. 7:30-8:30. Conference Opening. Welcome Catherine Snow, Harvard University. The Need for Research on Educational ...
www.educationaleffectiveness.org/conferences/2006/agenda.html - 12k -
Cached - Similar pages - Note this

Harvard University map — <10 km
hotels, restaurants, cafes, pubs near Harvard University. Nearest tags: cambridge lunch harvard square boston restaurant harvard food comics ...
www.tagzania.com/item/77797 - 51k - Cached - Similar pages - Note this

Figure 8. Proposed user interface for modern GIR systems

(2) A useful tool to aid users to better figure out their queries

From the other hand, a well-designed user interface of GIR could also aid users to figure out their queries, such as GeoVIBE system (Cai et al., 2002). Similar techniques are also applied in the field of SDI to build up a better search interface by incorporating the presentation of thumbnail map with the metadata information (Aditya and Kraak, 2007).

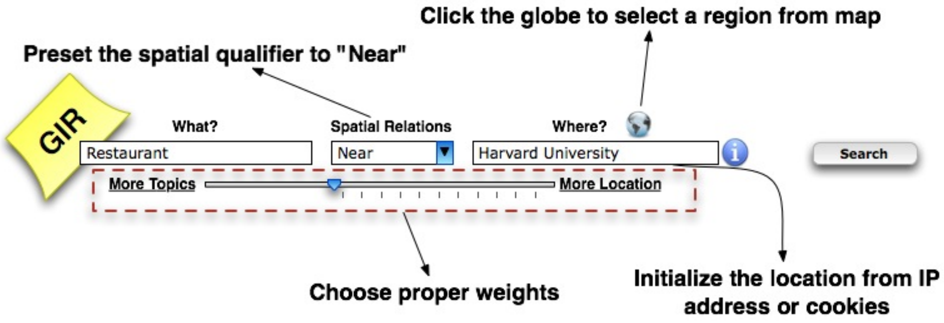
Another aspect of user interface design is automation of query parameters. For example, in most case, people are always interested in something of their locality or previous place he or she has searched for. Such information could be derived from the IP address of user's computer, which submits the query. Previous places or addresses could be saved as cookies in user's computer and restore before the next query. The spatial qualifier is always preset to "near" since it is the most frequent one that people use in their spatial query. The weights of thematic and geographical relevancy could also be set through the user interface (Fig. 10). Based on the optimized weights the GIR system provides from previous training set, user could choose the put more weight on the thematic component or the geographical component. It could be helpful if the user has a special information search need.

2.5 A new system architecture

As stated in the very beginning of this chapter, a new architecture needs to develop to integrate these key technologies that modern GIR system requires. The new architecture should also enable the modern GIR system to evolve by self-learning from users' feedback. All of these aspects require supports from the architecture level. More discussion about the new system architecture for modern GIR system will be addressed in the next section (Section III).

3. THE PROPOSED NEW FRAMEWORK FOR MODERN GIR SYSTEMS

As shown in the above figure (Fig. 11), the proposed framework of a GIR system is consisted of three main modules: (1) the user interface module, (2) the information extractor, storage and indexer module and (3) the query processing and information retrieval module.



Click the globe to select a region from map

Preset the spatial qualifier to "Near"

What? Restaurant

Spatial Relations Near

Where? Harvard University

Choose proper weights

Initialize the location from IP address or cookies

Figure 10. Timesaving automation of query parameters

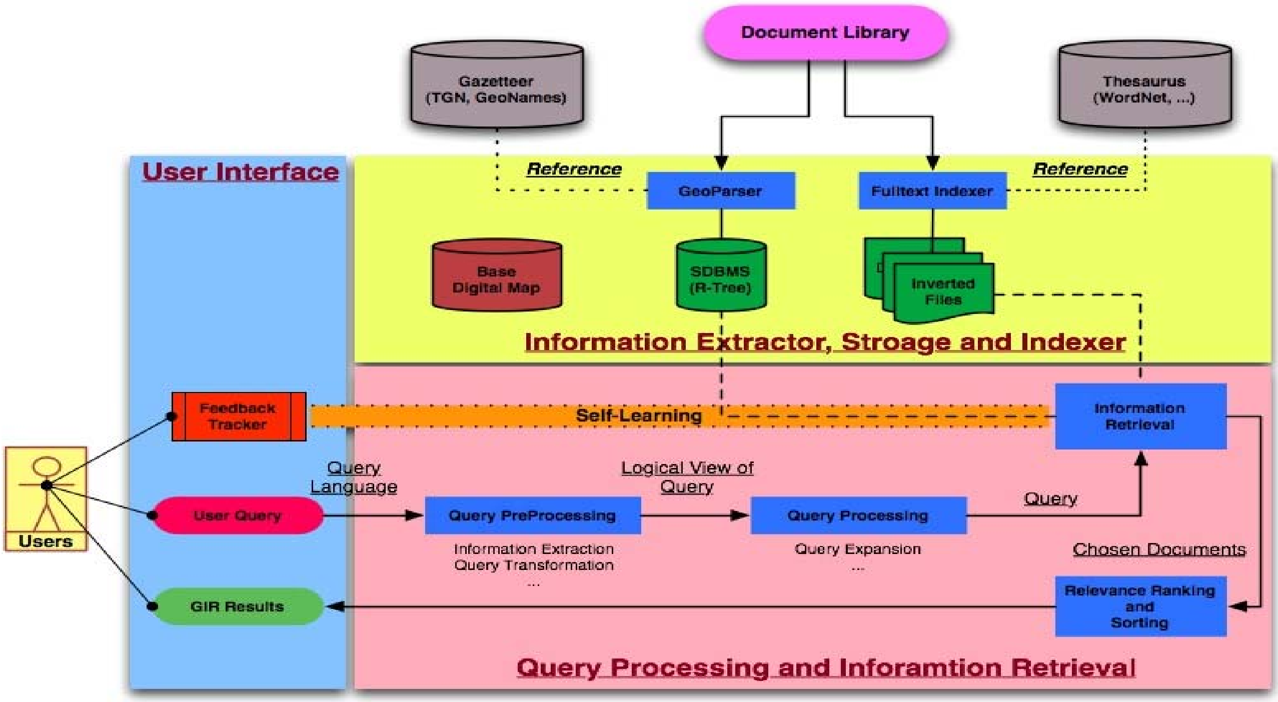


Figure 11. The proposed framework of modern GIR systems

(1) The User Interface Module

The user interface module will be in charge of the interaction between users and GIR system. It provides access points of users to specify and submit their spatial query, and then presents the result to users after that. Beyond the interaction for query collection and result presentation, there is one other important component inside the user interface module. It is the feedback tracker, which will anonymously and silently collect information about users' behaviour while using the GIR system. Users' judgment of documents' relevance could then de derived from their behaviour, and finally is used to adjust the information retrieval component. Such a GIR system with self-learning capability will become more and more advanced in retrieval accuracy after running for an enough long time.

(2) The Information Extractor, Storage and Indexer Module

There are two parsers/indexers, two data storage, one internal supporting data source and two external reference data sources.

- Two Parsers/Indexers: the GeoParser and Fulltext indexer. These two parsers/indexers will refer to the two external data sources: the Gazetteers (e.g. TGN, GeoNames, etc.) and thesaurus (e.g. WordNet, etc.).
- Two internal data storage: the SDB for storing the geographical footprints within text documents, and the inverted files for thematic information.
- One internal supporting data source: the digital map database, which will be used to create the background topographic map in the user interface.
- Two external reference data sources: the Gazetteer and Thesaurus, which will play an important role to match the toponyms and terms within text documents. The gazetteer also provides georeference to named places appeared in the documents.

(3) The Query Processing and Information Retrieval Module

The users' queries will be processed in this query processing and information retrieval module. There will be four components that make up this module.

- Query pre-processing component: inside this component, the related geographical and thematic information will be extracted from user's query. After that, the extracted information will then be delivered into the next step (query processing component).
- Query processing component: in this component, the user's query will be translated into the internal form, which could be processed directly by the information retrieval model. According to some heuristic rules and empirical study from training set, proper query expansion could be made to the original query. Hence, the GIR system could be able to retrieve documents according what the users really need, but not what he or she types in. The query expansion will also reduce the risk of losing some candidate documents from final result set.
- Information retrieval component: this component will retrieve relevant documents from document library, according to the similarity between them and the user's query. The query will be spited into two parts: the thematic part and geographical part. In the thematic part, conventional IR retrieval will be applied, while in the geographical part, the spatial query will be submitted and processed in the SDB. Then the two result sets will be merged to produce the final result set. The final result set of relevant documents will be sent to the ranking and sorting component before they are presented to the users.
- Relevance ranking and sorting component: relevance ranking of candidate documents will be carried out according their geographical and thematic closeness to the user's query. Before presenting the final results to end users, the documents will be sorted in a descend order according to their score of similarity. The most relevant document will receive the highest score and hence be presented on top of the result list.

4. CONCLUSION AND FUTURE WORK

In this paper, a new architecture of modern GIR system is proposed together with related key technologies to those problems the GIR community is facing nowadays. It is believed that the discussion and proposal presented in this paper could benefit the task of establishing a better GIR. But more practical works need to be done in the future. Improvement could be made to the proposed solution from the empirical study of practices.

REFERENCES

Cai et al., 2002. GeoVSM: An Integrated Retrieval Model for Geographic Information. Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg, pp. 65-79.

Christopher B. Jones and Ross S. Purves, 2008. Geographical Information Retrieval. International Journal of Geographical Information Science, 22(3), pp. 219-228.

Felix Mata, 2007. Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching. Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg, pp. 98-113.

H. Alani and C. Jones and D. Tudhope, 2001. Voronoi-based region approximation for geographical information retrieval with gazetteers. International Journal of Geographical Information Science, 15(4), pp. 287-306.

Purves R. S. et al., 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval. International Journal of Geographical Information Science, pp. 717-745.

Patricia Frontiera, Ray Larson, and John Radke, 2008. A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science, 22(3), pp. 337-360.

Simon Overell and Stefan Rüger, 2008. Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science, 22(3), pp. 265-287.

Trias Aditya and Menno-Jan Kraak, 2007. A Search Interface for an SDI: Implementation and Evaluation of Metadata Visualization Strategies. Transactions in GIS, 11(3), pp. 413–435.

Yu Liu, Yi Zhang, Yuan Tian and Lulu Xue, 2007a. On general place names and the associated ontology. Geography and Geo-Information Science, 23(6), pp. 1-7

Yu Liu, Yongxi Gong, Jing Zhang and Yong Gao, 2007b. Representation and Reasoning of Spatial Relations in Geographical Space. Geography and Geo-Information Science, 23(5), pp. 1-7

Baeze-Yates, R. and Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addition Wesley, Boston, MA

Charles T. Meadow, Bert R. Boyce, Donald H. Kraft and Carol Barry, 2007. Text Information Retrieval Systems (ed. 3rd). Academic Press, pp. 46-47.

Michael W. Berry and Murray Browne, 2005. Understanding Search Engines: Mathematical Modeling and Text Retrieval (2nd ed.). SIAM, pp. 21-22.

Bo Yu and Guoray Cai, 2007. A query-aware document ranking method for geographic information retrieval. *Proceedings of the 4th ACM workshop on Geographical information retrieval*, Lisbon, Portugal, pp. 49-54.

C. B. Jones et al., 2004. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Proceedings of Geographic Information Science: Third International Conference, Gi Science,* Adelphi, USA, pp. 125 - 139.

C. B. Jones et al., 2002. Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project. SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, ACM Press, pp.387 - 388.

C. B. Jones, Abdelmoty, and Gaihua Fu, 2003. Maintaining ontologies for geographical information retrieval on the web. *Proceedings of OTM Confederated International Conferences CoopIS.*

Gaihua Fu, C. B. Jones, and A. I. Abdelmoty, 2005. Building a geographical ontology for intelligent spatial search on the Web. *Proceedings of IASTED International Conference on Databases and Applications (DBA-2005)*, Innsbruck, Austria, pp. 167-172.

Xing Lin, Bo Yu and Yifang Ban, 2007. On Indexing Mechanism in Geographical Information Retrieval Systems. *Proceedings of 10th AGILE International Conference on Geographic Information Science,* Aalborg, Denmark, pp.1-3

Øyvind Vestavik, 2004. Geographical information retrieval: an overview. Dept. Computer and Information Science, Norwegian University of Technology and Science.