ON A NEW MODEL AND ALGORITHM TO ESTIMATE GEOGRAPHICAL SIMILARITY BETWEEN DOCUMENT AND QUERY IN GIR SYSTEM

Xing Lin

Division of Geoinformatics, Royal Institute of Technologies (KTH)
Drottning Kristinas väg 30, 100 44 Stockholm, Sweden
xingl@infra.kth.se

Commission WgS – PS: WG II/2

KEY WORDS: Spatial Information Science, GIS, Information, Service, Information Retrieval, Algorithm, Retrieval Model

ABSTRACT:

The usage of single geographical footprint model in existing Geographical Information Retrieval (GIR) system will cause many problems like overestimation or underestimation of the geographical scopes for documents to search. To be honest, the single geographical footprint model is not applicable in modern GIR system although it is simple, fast and widely applied today. In order to improve the quality of answers given to a spatial query, a new model as well as a dedicated algorithm will be proposed to study the geographical information attached to documents. Besides, a dedicated algorithm based on network graph, which is inspired by the Google PageRank and Bayesian network theory, is also invented to estimate the geographical similarity between document *d* and query *q* based on this new model. We believe that the new model and algorithm proposed here could better estimate the geographical similarity between document and query for a GIR system. Because it not only consider the geometric adjacency of document and query, but also take into account the importance of each places towards the hosting documents. By using the new model, the geographical scope of document could be better studied, and thus improve the quality of answers from GIR system concerning the spatial query.

1. INTRODUCTION

1.1 General Introduction

Most information in the world is linked to a certain place on the earth surface to some degree. Such information call always called geographical information. Geographical information exists in multiple forms such as cartographical maps, images, and texts (Cai, 2002). With the development of 3S (RS, GPS, GIS) techniques, more and more geographical information are collected and stored for various kinds of application. The WWW also hold vast amounts of information, most of which always have geographical footprint (Fu et al., 2005). Effective retrieval systems for geographical information are currently studied by both geo-spatial information scientists and library/information scientist (Cai, 2002).

By adopting the terminology of the SPIRIT project (Spatially-Aware Information Retrieval on the Internet), a typical geographically oriented search might looks like the following formula:

<*what, rel, where*>

which expresses the user's information search need of finding documents with the theme *what* which has a spatial relationship *rel* with the place or position indicated by *where* (Jones et al., 2002). A simple example of such spatial query might look like: "hotel near KTH", which means to find all the information about hotels which are near KTH.

Although modern web search engine, such as Google and Yahoo, have been widely used as the main tool of people to find information over the network or within organization (library, company, etc), conventional keyword-based search engines could not answer such spatial query well because they trust spatial terms involved in a query in the same way as other terms thus can not ensure good search results due to the lack of spatial awareness (Fu et al., 2005).

GIR (Geographical Information Retrieval) is the new promising theory and techniques to solve such problems. Recently, the fresh work of Semantic Web and Ontology have also been introduced into GIR theory to eliminate the vagueness and mismatch caused by different terminology used by information provider and the user making the spatial query. Answer a spatial query of <*what, rel, where*> using the primitive term from SPIRIT project in GIR system is based the estimation of similarity between the query and candidate documents from both the thematic aspect and the geographical aspect. The geographical information adhered to a document is always called geo-footprint. In most case, the single geo-footprint model or overall geo-footprint model is applied in existing GIR systems.

The usage of single geographical footprint model in existing GIR system will cause many problems like overestimation or underestimation of the geographical scopes for documents to search. As result of underestimation, not all documents related to the query could be returned. In the case of overestimation, all the related documents could be returned to the user but they might not be sorted in a correct order according to their adjacency to the query. To be honest, the single geographical footprint model is not applicable in modern GIR system although it is simple, fast and widely applied today.

1.2 Purpose and Structure of This Paper

In order to improve the quality of answers given to a spatial query, a new model as well as a dedicated algorithm will be proposed to study the geographical information attached to

documents. Besides, a dedicated algorithm based on network graph, which is inspired by the Google PageRank and Bayesian network theory, is also invented to estimate the geographical similarity between document d and query q based on this new model. We believe that the new model and algorithm proposed here could better estimate the geographical similarity between document and query for a GIR system. Because it not only consider the geometric adjacency of document and query, but also take into account the importance of each places towards the hosting documents. By using the new model, the geographical scope of document could be better studied, and thus improve the quality of answers from GIR system concerning the spatial query.

This paper will be divided into four sections. The first section will give a general introduction to the purpose of this paper. The problem of interest, as well as the basic idea of the author to solve this problem, will be described in this section. The second section is about the principle and theoretical analysis of the new model proposed in this paper to better estimate the geographical similarity between the documents and query in a GIR system. Several new concepts, such as geographical evidences, network of geographical evidences, etc., will be introduced in this section to help explain the principle of this new model. After that, the algorithm to calculate the geographical similarity using the proposed model will be given in Section 3 step by step. A small demo system is developed and introduced in the last section (Section 4). Based on the results of evaluation, some conclusion will be drawn about this new model. Improvements and future work will also be provided in this section too.

2. PRINCIPLE AND THEORETICAL ANALYSIS

2.1 Principle and Theoretical Analysis

From the experience of human spatial cognition, it is safe to make the assumption that a single document might contain more than one place names or mark. Usually these place names or landmarks contained in a single document are related to each other through a certain kind of spatial relationship (nearby/far, outside/inside, to the north/south/east/west of, etc.). It is also coincident with the human's intuition that if a place name A show up the most frequently in the while text of document, it is very possible that the focusing place this document is talking about should be A. While talking about the focusing place of A, there are also might other places, which have a close spatial relationship to the focusing place of A.

So it is could be concluded from the human's spatial cognition and intuition that:

(1) More than one place names are allowed to show up in a single document.

(2) The most frequent place name in the document is most likely to be the focusing place that this document is mainly talking about.

(3) Beyond the focusing place, there are other places in the same document. A network of place names could be established among the places within a document. In this network, the nodes denote the various places and they are connected to each other by the spatial relationship (in fact, the strongest spatial relationship) between them.

(4) The existence of places connected to a certain place A could be considered as evidences to reinforce or reduce the possibility that this document is talking about something happened in A.

(5) How the existence of a place influences the possibility of existence of another place in the same document depends on how strong spatial relationship exists between these two places. The strength of impact could be derived with the help of domain ontology of geographical information and spatial cognition.

(6) Ontology could be applied here to eliminate the problem of ambiguity caused by different terminologies. If all the places within a document are the same in the ontological view, which means they refer to the same place A on earth surface, the importance of this place against this document is 1 which means this document is mainly talking about something happened in this place A.

(7) If a given place B never shows up in the document and there are also no places in this document have a strong spatial relationship with the given place, the importance of this place against this document is 0. It means this document is not talking about something happened in this given place B at all.

(8) Based on the network of places within a single document, the score ranged from 0 to 1 could be calculated to estimate the importance value of how much degree this document is talking about something happened in a certain place.

(9) If a single geographical scope needs to be extracted, the most important one or the interpolated one could be proposed as the result. In case the interpolated approach is used, it will be a weighted average location by the importance of each place.

(10) Given a user query, the geographical adjacency is about the degree of matching between the geographical scope of query and document.

(11) Given the place A embedded within the user query, the geographical importance of a document upon the user search need is the importance of place A against the document.

(12) The final score of geographical similarity that incorporates importance and adjacency could be produced by multiplied the degree of adjacency with the value of importance.

Based on the above principles and analyses, we can better derive the geographical similarity between user's query and documents to match. Then this geographical similarity (denoted as $\mathrm{Re}l_G(q,d)$) could then be combined with similarity from the thematic scope (denoted as $\mathrm{Re}l_T(q,d)$) to generate an overall similarity score between documents and user's query. This overall similarity score could then be used to order the documents, which are considered by GIR systems to contain relevant information for the user information search need. The most relevant document will show up on the top and less relevant documents follow.

2.2 Important Concepts of Proposed New Model

The basic idea of this new model is *the concept of multiple geographical evidences instead of a single geographical footprint*. Each place in the document is served as an evidence to help determine which place this document is most possible to talk about or mainly talking about. The single geographical scope model is not applicable here. A score of geographical importance or probability from the places towards the document could be calculated to evaluate how spatially important places are towards documents. At the other hand, when considering about the geographical similarity, we also need to measure the geometric adjacency from the places of document to the geographical scope of query. A compound score of importance (denoted as $\mathrm{Re}l_G(q,d)$) from these two aspects could be

prompted to be the final geographical similarity between query and document.

The **geometric adjacency** $GA(q,d)$ could be estimated by those approaches that have been applied in the single geographical footprint model of existing GIR systems. For examples, the ratio of intersection and the Hausdorff distance (Greg, 2003) approaches are the applicable ones.

The **geographical importance** $GI(q,d)$ of places towards their hosting documents represents the probability of the fact that the hosting documents are really talking about something happened around them. Such importance has a great concern with the *frequency* of a certain place showing up in its hosting document, as well the ***context*** where each show-up takes place. The presence at the title place and emphasized region should weight more than that in the body text of articles. From the other hand, the *existence of other spatially related geographical entities* could also provide good evidences to study how much degree a single place is related to the document.

As inspired by the Google PageRank algorithm (Larry et al., 1998) and the Bayesian Network theory, **a network of geographical evidences** connected by spatial relationships will be established first.

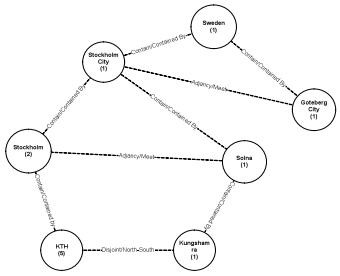

Figure 1. The network of geographical evidences within a document;

As shown in the above graph, we can investigate the relationships and inter-impact of geographical evidences within a single document. Such a network model will be of great help to estimate the degree of importance geographical evidence is towards the geographical topic of the hosting document.

When building up the network of geographical evidences for a selected document, many kinds of spatial relationship might be taken into account to connect nodes in the network. Usually, while estimating the importance towards the geographical topic of document, each kind of spatial relationship provides a different strength of impact from one geographical evidence upon the other connected ones. Not all the relationships between tow nodes will be showed up in the network graph. The spatial relationship will be examined and given a strength value of impact. Only the strongest spatial relationship will show up as link between two evidences in the graph. Not all connection between any two nodes will be drawn, but only those among siblings and between parent and child will be taken into account with the help of domain ontology as well as gazetteer/thesaurus.

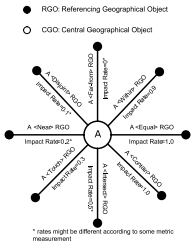

Figure 2. Impact strength of different spatial relationships

The strength of impact of spatial relationships connecting two geographical evidences is not easy to determine. Nevertheless, there are already some good attempts to examine the connectivity strength of spatial relationships in the viewpoint of spatial cognition. An adjacency model of topological relationships has been given by Bruns and Egenhofer (Bruns et al., 1996) to investigate the cognitive closeness between two geo-entities connected by different kinds of topological relationships. Another way to generate the strength of impact for each kind of spatial relationship could be achieved by carrying out a survey among potential users of the GIR system to see how they think about such impacts.

By studying the inter-impact as well as the frequency and context information using the algorithm proposed in this paper, the importance for each place within a document could be worked out with an acceptable quality.

2.3 The Proposed New Model

The geographical similarity between document d and query q could be calculated using the following equation (Equation 1).

$$\mathrm{Re}\,l_G(q,d) = \underset{i=0}{\overset{N-1}{Maximum}}(f(GA(q_G,d_{gi}),GI(d_{gi}))) \qquad (1)$$

where: f = the combination function of geometric adjacency and geographical probability;

N = the number of non-duplicated places within the document d;

gi = the i-th place of document d.

Possible combination functions could be the functions of addition or multiplication, as shown in the following equation (Equation 2).

$$f(GA(q_G,d_{gi}),GI(d_{gi})) = GA(q_G,d_{gi}) * GI(d_{gi}) \qquad (2)$$

Given the similarity from both the geographical and thematic scope, a compound score of relevance could be derived by combining these two similarity values. One of most popular way is the weighted linear combination as illustrated the following equation (Equation 3).

$$\mathrm{Re}\,l(q,d) = \omega_T * \mathrm{Re}\,l_T(q,d) + \omega_G * \mathrm{Re}\,l_G(q,d) \qquad (3)$$

where: $\mathrm{Re}\,l(q,d)$ = the overall similarity measurement between the document d and spatial query q;

ω_T = the weight of similarity measurement from the thematic scope;

$\mathrm{Re}\,l_T(q,d)$ = the similarity in the thematic scope;

ω_G = the weight of similarity measurement from the geographical scope;

$\mathrm{Re}\,l_G(q,d)$ = the similarity in the geographical scope;

3. METHODOLOGY AND ALGORITHM

The most significant advantage of the new model proposed in this paper against the existing ones is introducing the concept of multiple geographical evidences within a document and considering the geographical importance of each place to the hosting document when calculating the geographical similarity between the user's query and documents to retrieve. The algorithm to estimate the geographical importance of places towards their hosting document could be expressed in the following steps.

Step 1: Extraction of geographical evidences and Building up the network

After the lexical analysis and work splitting process of documents, extract the place names and address text as geographical evidences with the help of domain ontology of geographical information and gazetteer/thesaurus. The network could then be built up using the geographical evidences as nodes and the spatial relationships within as links.

Step 2: Initial distribution of importance based on the frequency weighted by context

Setting the maximum of importance value as M, the initial importance GI_i^0 of geographical evidence g_i could be derived using the following formula (Equation 4):

$$GI_i^0 = \frac{M}{\underset{m=0}{\overset{N-1}{Maximum}}\left(\sum_{n=0}^{freq(m)-1}\omega_{m,n}\right)} * \sum_{n=0}^{freq(i)-1}\omega_{i,n} \qquad (4)$$

where: GI_i^0 = the initial importance for the geographical evidence g_i;

$freq(m)$, $freq(i)$ = functions to calculate the frequency of the given geographical evidence;

N = the total number of places (not duplicated) within the documents;

$\omega_{m,n}$ = the weight for the m-th geographical evidence at the n-th occurrence. This weight is determined by the context where this place shows up.

Step 3: Iterative procedure based on network structure

Given on the initial importance, an iterative procedure is carried based until it runs enough rounds or meets a stop indicator. For each round before it stops, the procedure could be divided into 3 steps.

Step 3.1: Basing on the network structure, incorporate the impacts of neighbours with their importance values from last step. The most direct and simplest approach is shown in the following section, which achieves its goal by summing up the weighted impact of neighbours together with the original impact. Given a geographical evidence g_i to evaluate the importance, suppose $S_i = \{R_i \mid 0 \le i \le N_i - 1\}$ as the set of all the directly connected geographical evidences in the network; N_i is the number of referencing geographical evidences spatially connected to g_i. Then the incorporated importance of g_i after this round could be estimated using the following formula (Equation 5):

$$GI_i'^{j} = GI_i^{j-1} + \sum_{m=0}^{N_i-1}(GI_{R_m}^{j-1} * I_{i,R_m}) \qquad (5)$$

where: j = the round number;

$GI_i'^{j}$ = the j-th importance value of g_i before normalization;

GI_i^{j-1} = the importance value of g_i from last round;

$GI_{R_m}^{j-1}$ = the last importance value of reference object R_m;

I_{i,R_m} = the impact index between g_i and R_m caused by the spatial relationship between them.

Step 3.2: All $GI_i'^{j}$ will be normalized and resize to be within the range of $[0 \quad M]$ before going to the next round using the Equation 6.

$$GI_i^j = \frac{M}{\underset{m=0}{\overset{N-1}{Maximum}}(GI_m'^{j})} * (GI_i'^{j}) \qquad (6)$$

Step 3.3: Update the network status and repeat the Step 3.1 and Step 3.2 until enough rounds have been run or the system meets the threshold to stop.

After these processes, all the places contained in a document will gain an index of GI_i geographical importance towards their hosting document. Such geographical importance values will range from 0 to M.

$$\underset{i=0}{\overset{N-1}{Maximum}}(GI_i) = M \quad and \quad \underset{i=0}{\overset{N-1}{Minimum}}(GI_i) = 0 \qquad (7)$$

4. EVALUATION AND CONCLUSION

4.1 Demo System and Evaluation

In this paper, a small demo system is developed in order to evaluate how well the proposed model acts in answering the users' query with geographical semantic. This experimental work will mainly care about the accuracy by examining how retrieved documents meet the information search need of GIR system users. Given a spatial query submitted to the demo system, if all the geographically relevant documents are returned in the results and more relevant documents have a higher score of similarity, it could be concluded that the proposed model performs well in answering users' questions from the viewpoint of geographical aspect.

Due to limitation of sample dataset size, functionalities and purpose of this demo system, we will only take care the geographical similarity between documents in library and the user's query. Moreover, the real score of relevance will be judged by manually interpreting the content of documents to see how similar it is as the user's query in the geographical aspect.

The demo system in this paper is built based on the Apache Lucene (Lucene, 2007), the most famous open-source search engine toolkit, and the JTS Topology Suite (JTS, 2007), which provides Java implementation of OGC specification of simple feature (OGC, 2006) and bundles of useful geometric algorithms.

Regarding the sample documents used in this experiment, about 100 articles of US travel guides are randomly collected from the WorldWeb (WorldWeb, 2007) website. These articles are mainly talking about some touring sites all through the United States. Besides, a small gazetteer is established from the boundaries and names of US states, counties and cities are extracted from the ESRI Data & Maps (ESRI, 2007) shipped together with ArcGIS software package. This small gazetteer is used to help extract places from the text of sample documents.

The sample documents are first parsed and indexed by the Lucene toolkit. The JTS is then used to determine the spatial relationships between geo-entities. The geographical adjacency between geo-entities is measured as the ratio of intersection and is also computed by using the JTS package.

To submit a spatial query to this demo GIR system by specifying the topic and geographical location, such as "Pub in Washington", the system will return a list of thematically relevant documents, which are ordered by the geographical similarity. Here the combination of thematic and geographical

similarity is ignored because the combination of these two could be as complex as the whole content of this paper. The following image (Figure 3) is an example of a query and results returned.

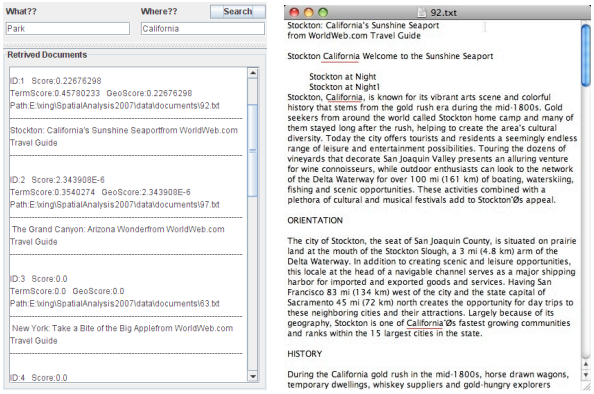

Figure 3. Demo system: interface and results.

By manually interpreting the text of each retrieved document, it could be found that the top documents are always the most geographically relevant with the user's query.

4.2 Conclusion, Limitation and Future Work

From the results of evaluation, it could be found that the model performs well in measuring the geographical similarity between user's query and documents to retrieve. Comparing to the single geographical footprint model or the overall geographical footprint model, the method proposed in this paper considers both the geographical adjacency and the geographical importance of a place on earth against its hosting document. Based on the concepts of multiplicity and network of geographical evidences, the model proposed in this paper is more natural and more accordant with the common sense of human cognition in the geographical domain. As a result, the compound score of geographical similarity could better represent the degree of relevance between the document d and query q in the geographical aspect.

Obviously, this model is not a perfect one. Although this paper provides an approach to calculate the geographical importance of a place against its hosting document, the combination of geographical importance and geographical adjacency to get the geographical similarity demands further study. In this paper, the most simplest combination method of multiplication is adopted. The strength of impact of different kinds of spatial relationships also demands more attention. It makes the calculation of impact strength more difficult when the cognitive closeness outweighs the Euclidian distance. What's more, the success of this new model relies greatly on the correct extraction of geographical places from the text of documents. With the growing of number of geographical evidences within a single document, the network might be too complex that it takes a long time to calculate the geographical similarity for a given query. Bigger testing library of documents should be built to make a full evaluation of the real performance of this new model.

Another question worthy of further consideration is the combination of geographical and thematic relevance score. Although these two are usually merged using the formula in Equation 3, how to determine the values of ω_T and ω_G remains to be an open question. It is because sometimes people care more about the thematic topic, while in other case the geographical location means more than the thematic topic. Recently, Yu and Cai (Yu et al, 2007) developed an approach to dynamically determine these two weights by analyzing the specificity of user's query. Although their approach is a good attempt to this question, more evaluation and improvement could be applied to their method.

References from Journals:
(Empty)

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

References from Other Literature:
Bruns H. T., et al., 1996. Similarity of Spatial Scenes. *Proceedings of the 7th International Symposium on Spatial Data Handling*, Delft, The Netherlands, pp. 173-184.

Fu et al., 2005. Building a Geographical Ontology for Intelligent Spatial Search on the Web. *In Proceedings of IASTED International Conference on Databases and Applications (DBA-2005),* Innsbruck, Austria, Springer Verlag, pp. 167-172.

Jones et al., 2002. Spatial information retrieval and geographical ontologies: an overview of the SPIRIT project. *Proceedings of the 25th annual international ACM SIGIR*, Tampere, Finland, pp. 387-388.

Larry P., et al., 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University.

Yu et al., 2007. A query-aware document ranking method for geographic information retrieval. *Proceedings of the 4th ACM workshop on Geographical information retrieval*, Lisbon, Portugal, SESSION: Query and results processing, pp 49-54.

References from websites:
ESRI, 2007. ESRI Data & Maps (version 9.2). http://www.esri.com/data/index.html (accessed 29 Jan. 2008)

Greg J., 2003. Spatial similarity functions, http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html (accessed 29 Jan. 2008)

JTS, 2007. JTS Topology Suite Website. http://www.vividsolutions.com/jts/jtshome.htm (accessed 29 Jan. 2008)

Lucene, 2007. Apache Lucene Project Website. http://lucene.apache.org/ (accessed 29 Jan. 2008)

OGC, 2006. OpenGIS Simple Feature Specification for SQL. http://www.opengeospatial.org/standards/sfs (accessed 29 Jan. 2008)

WorldWeb, 2007. United States Travel & Tourism. http://www.usa.worldweb.com/ (accessed 29 Jan. 2008)