Relevance and Ranking in Geographic Information Retrieval

Chandan Kumar
OFFIS - Institute for Information Technology,
Oldenburg, Germany
chandan.kumar@offis.de

Geographic Information Retrieval (GIR) is a specialized branch of traditional Information Retrieval (IR), which deals with the information related to geographic locations. One of the main challenges of GIR is to quantify the spatial relevance of documents and generate a pertinent ranking of the results according to the spatial information needs of user. Most of the current methods judge the relevance of documents just based on textual and spatial similarity with the query, and ranked the results with a linear combination of these similarity measures. We consider relevance ranking as a much more dynamic problem stemming from real world application such as location based mobile services, where user not only seek information but there is a decision making involved with the search i.e. to visit the location. In this paper we discuss current ranking phenomenon in geographic information retrieval, present different relevant parameters based on our initial study, and argue for the need of a formal relevance framework and ranking mechanism for geographical information retrieval. We approach GIR ranking as a spatial decision problem to support user's activity, and propose the idea to explore decision-theoretic framework and probabilistic representation for geo relevance formalization.

Geographic information retrieval, Location based search, Geo-relevance, Decision theory

1. INTRODUCTION

Information Retrieval is a discipline which deals with the storage, organization, representation and access of information. The goal of an IR system is to return relevant documents in response to a user query. The term relevant means that retrieved documents should be related to the user's current information requirement. The problem of estimating the document relevance is usually regarded as a ranking problem to present the ordered list of documents related to a query. In traditional IR ranking models, the effectiveness of results depends on the appropriateness of a user query, and the performance is limited when desired results have inherent information requirements. This problem especially applies on geographic Information retrieval which is a specialized branch of IR that deals with the retrieval of documents with geographic significance. Here all the georeferenced documents represent real world physical objects, and approaching information based on their georeferences can be important in several contexts, and the relevance judgment requires interpretation of implicit information enclosed in documents and queries to provide suitable response to queries.

Relevance formalization in GIR is still a new field of research that has so far been vaguely defined. The current work in geographic relevance ranking has simplified the problem of judging the relevance of a document to the problem of combining the textual and geographical similarity between the query and document. While typical information retrieval measures have been explored to compute thematic importance, geographical importance has been reduced to computing the similarity between two geographic locations, one associated with the query and the other with the document. These hybrid ranking methods do not consider the document specific spatial properties and its connectivity to other documents and are less adaptive to spatial and contextual information need especially for a user in mobile environment. In the proposed research we aim to study the geographic information behavior of users to argue various geo relevance parameters. Our objective is to establish a formal ranking mechanism for GIR, including all the relevance parameters to quantify the spatial properties of documents and provide significant contextual results.

The remainder of this paper is organized as follows. Section 2 briefly describes GIR architecture and

current method for relevance ranking. In section 3 we present the different relevance parameters to judge the significance documents in our proposed GIR ranking framework. In section 4 we introduce the idea of probabilistic representation and decision theoretic framework as a formal approach to GIR ranking problem, and at the end section 5 concludes the paper.

2. GEOGRAPHIC INFORMATION RETRIEVAL

For many of our daily activities geographic locations and geographic entities play an important role - where we are, where things and peoples are, where to get there, what can be found in some area and so on. Geographic search is especially important for location-based services where a user in a mobile environment might have dynamic and contextual information demands. Geographic Information retrieval is a new and growing field of research which supports the search of geographic preferences. The goal of GIR systems is to process the information request and provide the results to satisfy spatial information need.

The major challenges in GIR systems are: analyzing and processing the document collection and queries, creating textual-geographical indexing, and ranking of the document using a particular relevance mechanism. The first two problems are normally seen as necessary pre-processing tasks, so that the latter one can use ranking formulas between the documents and of user queries. Preprocessing includes the methods such as information extraction of geographic terms from structured and unstructured data; ontology creation, word sense disambiguation, geographic phrase translation etc. The task is to first build a spatial document corpus, then analyzing the corpus, identifying references to locations, disambiguating them and building a geographic index. Previous research in GIR has addressed these problems of recognition and disambiguation of place references given over text with methods Leidner (2007); Martins et al. (2010), and the assignment of documents to encompassing geographic scopes Anastacio et al. (2009). While there has been lot of research efforts on geoparsing, geocoding, and geospatial crawler's Ahlers and Boll (2009, 2008, 2007) techniques to build a high-quality geo database, ranking of documents in response to a user query is still a challenging research problem. Figure 1 explains the basic architecture of a GIR system. GIR has also been evaluated at the CLEF forum since year 2005 as GeoCLEF task Mandl et al. (2008). The results show that traditional IR ranking methods are able to retrieve many relevant documents for most spatial queries, but it remains difficult to generate a significant ranking of them.

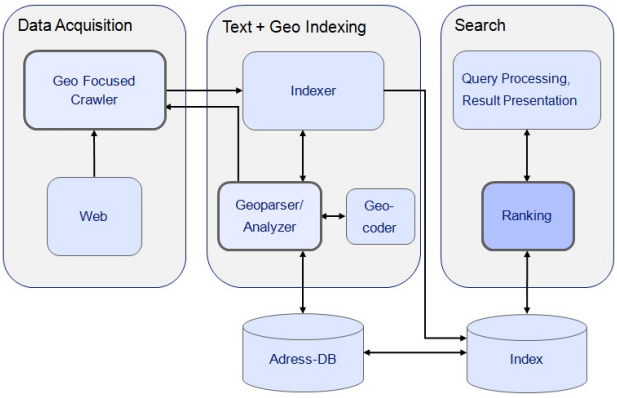

Figure 1: *Geographic Information Retrieval Architecture*

Due to this evidence, recent GIR approaches have especially focused on the ranking mechanisms.

2.1. Current Ranking Methodology in GIR

An essential part of GIR system is to assign a relevance score to documents to represent how well they fulfill a user's spatial information need. In general, both the documents and geographic queries can be expressed as lists of extracted location names as well as their original texts. The current work in geographic relevance ranking has simplified the problem of judging the relevance of a document by splitting the thematic and geographic relevance estimation, with the assumption that they are independent from each other. While typical information retrieval measures have been explored to compute thematic relevance (query-document thematic similarity), geographical relevance is the closeness between two geographic locations, one associated with the query and other with the document. There have been different methods to find the geographic relevance with different similarity methods. Frontiera et al. (2008) compared different similarity measures based on region overlaps. Martins et al. (2007) proposed a similarity function for GIR that, instead of using area overlaps, uses a non-linear normalization of the distance between the document and query scopes. Henrich and Ludecke (2009) noticed that the exact similarity between the regions is not the main focus in GIR, they found out that overlaps are just a strict notion of similarity (e.g., two regions that are near each other but not overlapping are just as dissimilar as two regions that are hundreds of miles apart) and, for GIR, similarity metrics should also account with other perspectives besides overlap. Larson and Frontiera (2004) presented a logistic regression model to estimate the spatial overlap between the query and document.

After the independent computation of thematic and geographic relevance the ranking problem is to

combine these two types of relevance. There have been various schemes to combine two types of relevance scores but the most common one is the weighted sum of individual scores Martins et al. (2005); Cai (2002); Andrade and Silva (2006).

$$Rel(q,d) = W_T * Rel_T(q,d) + W_G * Rel_G(q,d) \quad (1)$$

Where q is a query, d is a document. Rel_T Rel_G are functions to calculate the textual and geographic relevance. W_T and W_G are weights of these two individual relevance scores respectively. The traditional vector space model of IR has been extended to a hybrid geothematic IR model to include the geographic dimension of relevance, called GeoVSM Cai (2002). It measures document relevance in geographical and thematic subspaces differently, producing two scores respectively. It uses the above mentioned weighting scheme to combine the relevance. Other variations of the VSM model such as random indexing (RI) which accumulates context vectors for words based on co-occurrence data has also been explored for GIR ranking Carrillo et al. (2010).

Martins et al. (2005) proposed various other combination functions such as the product or the maximum of two individual scores to compute the final ranking score. To achieve some kind of dynamism in ranking, Yu and Cai (2007) proposed a dynamic document ranking scheme to combine the thematic and geographic relevance measures on a per-query basis. The authors used query specificity (i.e., the geographic area covered by the query) to determine the relative weights of different sources of ranking evidence for each query (i.e., the weight of the geographic relevance measure is inversely proportional to the area of the query). Though the approach provides some kind of flexibility to the end results, its limited to per query analysis and still follows the same hybrid relevance combination approach. Martins and Calado (2010) describes the usage of a SVM learning to rank approach to combine different metrics of thematic and geographic similarity into a single ranking function. Like other approaches this again considers only the query dependent parameters, still it provides a good alternative to traditional static combination methods, but it depends upon the feature selection mechanism on training dataset for ranking, which is still an unsolved problem and in need of more research efforts.

In comparison to computational methods to estimate ranking score, there are other approaches which focus on visualization of multiple dimensions of relevance to assist human judgment of relevance, Hobona et al. (2006). These methods assumed that the end users themselves must be the ultimate judges to the relevance of documents, and the ranking algorithms cannot provide the final judgment on relevance. So, they facilitate the users to find relevant documents with multidimensional visualization mechanisms to represent the degrees of relevance. Cai (2001) presented GeoVIBE to explore the possibilities of visualizing documents in both geographic and thematic domains. The system has two browsing windows, GeoView and VibeView, to display geographic and thematic relevance. Hobona et al. (2006) proposed a three-dimensional visualization environment to present three degrees of relevance In their model. They used 3D vector space to represent the geographic, thematic, and temporal similarities of a geospatial dataset.

The visualization approaches offer more freedom for users to express their information needs and make own decisions on relevance, but these sophisticated multidimensional interfaces cause extra cognitive load and require lot of learning efforts. More importantly automated ranked list of documents fulfill end users current search habits. Thus, we focus on the more realistic latter approach of listing result documents with computational ranking solutions.

3. RELEVANCE PARAMETERS FOR RANKING

Document relevance in GIR has been mostly judged by thematic and geographic similarity measures but there are many other aspects which could affect the rank of a document. This could be understood with a simple example query like ‘Museums near Munich’, there could be several documents about the same museum. The nearby museums would be ranked high, However, the user probably also wants documents about other museums that may be a bit further away, especially when these documents are more relevant for the term ‘museum’. So the ranking mechanism should be an ideal way to combine the content and geographic relevance of documents with respect to the user query to take care of relevance, redundancy and novelty. There could be other important factors which are independent of the query but significant to provide better results. For example a museum which is close to other museums might be more interesting for user (relation between location objects), or the number of users visited the location shows the significance of a particular object compared to others. There could be several user dependent factors as well to change the order of the results. A user interested in arts and painting, would prefer to see museums of modern art and sculpture. A user in mobile environment might have some time constraints, or the museum located in the driving direction? or the significance of distance and thematic relevance might be different for a automobile user, a pedestrian etc. There could be

many such factors to rank a document, thus, we categorize such parameters in the following four major relevance categories.

- **Query biased thematic relevance:** This relevance measure is to estimate the aboutness of a document with respect to query topic, or the thematic match between a textual content of a document and given query. It's a general phenomenon in information retrieval settings and research in GIR has acquired similar methods to calculate the thematic relevance.
- **Query specific location relevance:** This concerns geographic dimensions of relevance when matching documents to queries, here relevance is judged by the spatial relationships (such as contain, overlap, intersect, connect, near etc.) between documents and queries in geographical space. It is common to use distance functions between the query and document footprint to estimate geo relevance.
- **Query independent spatial relevance:** Apart from the query-dependent relevance discussed above the aim is to utilize the static query-independent relevance such as link structure of the Web. Similar to the PageRank algorithm that models popularity and importance by the number of incoming links of a page, a link-based method that takes geospatial properties of Web pages into account should provide a better measure of geospatial importance. The motivation is the need for a quality indicator of geospatial search results that is independent of a single page. While an individual page may carry an address and contain rich content, it lacks context information that allows assessing its credibility. The incoming links of a page should not be used homogeneously, but given a differentiated weight according to the geographic distribution and spatial characteristics of the source of the incoming links.
- **User centered contextual relevance:** Current geographic information retrieval ranking methods are mainly concerned with retrieving documents from the location perspective, such as user's current location as the only geo relevance criteria. But approaching information based on their georeferences can be important in several user centered contexts. There are many factors such as temporal constraints, environmental circumstances, current activity, personal interest and preferences of a user that could influence the importance of a page for the spatial information need. The goal is to characterize the user's implicit or explicit information need for geographic significance.

Geographic context relevance is the topic discussed in various geo community research articles. Raper Raper (2007) formally proposed the concept of geographic relevance considering Information seeking issues of mobility and geography. More recent work on geo relevance Sabbata (2010); Sabbata and Reichenbacher (2010) discusses various criteria of geo relevance, they do not reference the ranking problem of GIR or its impact on Web document-query relationship, but the conceptual modeling of spatial objects and its relation to user's context and activity present various interesting issues, and motivate us to utilize spatial context information for actual ranking in GIR.

4. FORMALIZATION OF RELEVANCE PARAMETERS FOR RANKING

The proposed research is focused towards defining each of the parameters described above and exploring their significance on geographical context. In doing so, a single ranking framework is required to optimize all the measurable evidence. While most of current ranking methods in GIR have employed a static way to combine the thematic and geographic relevance, with no concern about the different contexts of users' search behavior, our research objective is the formal modeling of all the relevance parameters and present an effective ranking mechanism based on a proper theoretical framework. We aim to formalize the ranking as spatial decision problem, as we believe that the goal or GIR system is to assist real time decision making of end users. Geographic information retrieval actually attempts to provide intellectual and physical access to geo-referenced information sources, i.e., the relevance in geographic retrieval is closely attached to the physical world, than to the informational world to satisfy contextual information need and support user activity. So the ranking of document in geographic information retrieval is a spatial problem to solve, to support user's activity and decision making. Decision theoretic framework is a natural probabilistic phenomenon to support user activity Berger (1985). In geographic information retrieval system each document represents a certain location, or possible destination that user is searching for, so as per decision theoretic terms the document collection could be an action space, where each action represents a document. Each possible action should be evaluated based on all available parameters (discussed above) associated with it and the best possible actions would be ranked higher. There can be many possible criteria to choose a particular action, e.g., the risk/error minimization principle has been quite successful for information science studies

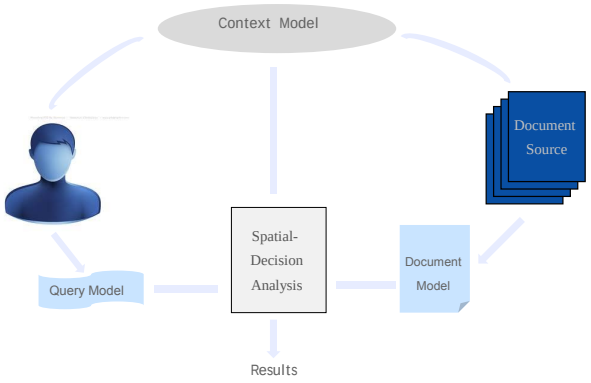

Figure 2: *probable ranking framework*

due to its strong information theory and probabilistic background Kumar et al. (2009); Lafferty and Zhai (2001); Lavrenko and Croft (2001).

We aim to explore the probabilistic representation scheme for modeling purpose, as given the imprecise and incomplete way in which a user's information need is represented by a query and a relevant document by its indexing, relevance should be approached probabilistically. This is the argument given by Maron and Kuhns (1960) to introduce probabilistic framework in information retrieval, which is especially true in geographic information retrieval since all the information objects are abstract, compressed representation of real world phenomena and that contains some degree of error and uncertainty Goodchild (1999), also the empirical studies on real Web search logs by Backstrom et al. (2008) shows that the spatial distribution of query issuers to a specific event is highly variant. So the goal is to model query, document, and context probabilistically for the ranking purpose. The effectiveness of probabilistic measures such as BM25 and logistic regression Larson and Frontiera (2004); Martins and Calado (2010) for spatial similarity computation also support our idea to use probabilistic modeling for geo ranking. Figure 2 shows our roadmap to develop a ranking mechanism with models query, document and context to solve the spatial decision problem of end users.

5. CONCLUSION

In this paper we discussed the research problem of relevance ranking in geographic information retrieval. Most of the current methods use the textual and spatial similarity as only relevance criteria, and employed a static way to combine them which is less flexible and adaptive to search contexts. This type of ranking solutions is difficult to be tuned and optimized, and the distribution of thematic relevance and spatial relevance lacks necessary justification. We consider that the problem of geo relevance ranking is much more complex, and current approaches offer only partial solutions. Our research goal is to use different relevance parameters including document's prior spatial importance and contextual significance to judge the importance of a document. We plan to approach geographic ranking as a spatial decision problem, and model all the parameters in a single probabilistic relevance framework.

6. REFERENCES

Ahlers, D. and Boll, S. (2007) *Location-based Web search* In A. Scharl and K. Tochtermann, editors, The Geospatial Web. Springer.

Ahlers, D. and Boll, S. (2008) *Retrieving Address-based Locations from the Web* In C. Jones and R. Purves, editors, GIR'08. ACM.

Ahlers, D. and Boll, S. (2009) *Adaptive geospatially focused crawling* In Proceedings of the 18th Conference on Information and Knowledge Management.

Kumar, C., Pingali, P. and Varma V. (2009) *Estimating Risk of Picking a Sentence for Document Summarization* In 10th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-09, Mexico City, Mexico.

Andrade, L and Silva, M. J. (2006) *Relevance Ranking for Geographic IR* Proceedings of the workshop on Geographic Information Retrieval, SIGIR 06, Seattle, USA.

Raper, J. (2007) *Geographic relevance* Journal of Documentation, 63(6):836-852.

Sabbata, S.D. (2010) *Criteria of Geographic Relevance* Proceedings of the 6th International Conference on Geographic Information Science, Zurich, Switzerland.

Sabbata, S.D. and Reichenbacher, T. *A probabilistic model of geographic relevance* Proceedings of the 6th Workshop on Geographic Information Retrieval, February 18-19, Zurich, Switzerland.

Goodchild, M. (1999) *Future directions in geographic information science* Geographic Information Science, 5(1), 1-8.

Maron, M., and Kuhns, J., (1960) *On relevance, probabilistic indexing and information retrieval* In Journal of ACM ,216-244

Cai, G. (2002) *GeoVSM: An Integrated Retrieval Model For Geographical Information* Lecture Notes on Computer Science 2478: Geographical Information Science: Second International Conference on GIScience, Baltimore, MD, USA, 2002, 65-79.

Martins, B., Silva, M. J., and Andrade, L. (2005) *Indexing and Ranking in Geo-IR Systems* Proceedings of the workshop on Geographic Information Retrieval, CIKM 05, Bremen, Germany.

Martins, B., Calado, P.(2010) *Learning to rank for geographic information retrieval* In: Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR '10, ACM.

Lafferty, J. and Zhai, C. (2001) *Document Language Models, Query Models, and Risk Minimization for Information Retrieval* In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.111-119, New Orleans, Louisiana, United States.

Lavrenko, V. and Croft, W. B. (2001) *Relevance based language models* In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, p.120-127, New Orleans, Louisiana, United States.

Martins, B., Cardoso, N., Chaves, M.S., Andrade, L., and Silva, M.J. (2007) *The University of Lisbon at GeoCLEF 2006* In Proceedings of the 6th Cross-Language Evaluation Forum.

Yu, B. and Cai, G. (2007) *A query-aware document ranking method for geographic information retrieval* In Proceedings of the 4th ACM Workshop on Geographical Information Retrieval.

Henrich, A. and Ludecke V. (2009) *Measuring Similarity of Geographic Regions for Geographic Information Retrieval* In Proceedings of the 31st European Conference on Information Retrieval.

M. Carrillo, E. Villatoro, A. Lopez, C. Eliasmith, L. Pineda, M. Gomez. *Concept Based Representations for Rankingin Geographic Information Retrieval* In Proceedings of the 7th international conference on Advances in natural language processing.

Berger, J. (1985) *Statistical decision theory and Bayesian analysis* Springer, Heidelberg.

T. Mandl, F. Gey, G. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, and X. Xie. (2008) *GeoCLEF 2007: the cross-language geographic information retrieval track overview* In Working Notes for the Cross Language Evaluation Forum Workshop.

J. L. Leidner (2007) *Toponym Resolution in Text* PhD thesis, University of Edinburgh.

I. Anastacio, B. Martins, and P. Calado (2009) *A Comparison of Different Approaches for Assigning Documents to Geographic Scopes* In Proceedings of InForum 2009, the 1st Portuguese Symposium on Informatics.

B. Martins, I. Anastacio, and P. Calado (2010) *A Machine Learning Approach for Resolving Place References in Text* In Proceedings of AGILE-2010, the 13th AGILE International Conference on Geographical Information Science.

P. Frontiera, R. Larson, and J. Radke (2008) *A comparison of geometric approaches to assessing spatial similarity for GIR* International Journal of Geographical Information Science.

Larson, R. and Frontiera, P. (2004) *Ranking and Representation for Geographic Information Retrieval* SIGIR 2004 Workshop on Geographic Information Retrieval, Sheffield, UK.

L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. (2008) *Spatial variation in search engine queries* In Proc. of WWW 2008, pages 357-366, New York, NY, USA.

Cai, G. (2001) *GeoVIBE: A Visual Interface to Geographic Digital Library* Proceedings of the 1st Visual Interfaces to Digital Libraries Workshop, Roanoke, VA, USA, 2001.

Hobona, G., James, P. and Fairbairn, D. (2006) *Multidimensional Visualization of Degrees of Relevance of Geographic Data* International journal of geographic information science, 20(5), 469-490.