A Probabilistic Model of Geographic Relevance

Stefano De Sabbata
University of Zürich
Winterthurerstr. 190
8057 Zürich, Switzerland
stefano.desabbata@geo.uzh.ch

Tumasch Reichenbacher
University of Zürich
Winterthurerstr. 190
8057 Zürich, Switzerland
tumasch.reichenbacher@geo.uzh.ch

ABSTRACT

In this paper, we present a new model for the assessment of Geographic Relevance. This model is drawn from Okapi BM25, thus it takes into account not only a score for each dimension of relevance but also the distribution of these scores within the collection. Preliminary results suggest that the relevance estimation of top-ranked objects is more sensitive to small changes in the user context.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms

Theory, Algorithms

Keywords

Geographic Relevance, Okapi BM25, GRBM25

1. INTRODUCTION

The concept of Geographic Relevance (GR) was first proposed by Raper in [7]. GR can be defined as a relationship between the user's geographic information needs and the spatio-temporal expression of geographic objects in the user's surrounding environment.

In [8] the authors propose a conceptual framework of GR, that includes a definition of the user context and a definition of the geographic objects. Both these definitions are composed by elements that describe the related entity in the dimensions of space, time, concepts, etc. The purpose is to analyse the relationships between the elements describing the user context and the elements describing the geographic objects, where each relationship is a facet of GR in a different dimension (i.e. space, time, concepts, etc.). The assumption is that a distance value can be measured for each relationship in the related dimesion, that is to quantify GR with this set of distance values.

To estimate the relevance of a geographic object for a given user context, these distance values have to be converted to similarity scores. A possible approach is to deal with this conversion in the same way the Geographic Information Retrieval (GIR) systems deal with the spatial dimension. Most GIR methods take into account two different types of similarity between a user defined query and a document: a textual similarity and a spatial similarity. The first is often calculated by the Okapi BM25 model [3]; the latter is usually calculated as a function inversely proportional to the distance between two locations. Examples of this approach can be found in [1, 6]. Likewise, the approach proposed in [4] can be seen as the use of binary functions on different relevance dimensions (e.g. space, time, visibility).

Still, these approaches do not deal with the distribution of the objects in the user's surrounding environment. For instance, applying one of these approaches to the spatial dimension, an object would have the same similarity score for a given distance, no matter if it is the closest object to the user or if there are a large number of closer objects.

To overcome the above-mentioned issue, we propose a new model drawn from Okapi BM25 and thus named GRBM25. This model can be applied to any generic relevance dimension, given a distance function and a related inversely proportional function. In particular, our aim is to assess the GR of a geographic object for a given user context, applying this model to a set of relevance dimensions, defined by a related set of relationships between the two entities described by the GR framework.

2. MODEL FORMALIZATION

Let $c \in C$ be a user context description, $G = \{g_1, g_2, \dots\}$ a set of geographic objects, and $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ a set of relationships, between $c \in C$ and $g \in G$, defined within the GR framework. Afterward, we assume that:

$$\forall \gamma_i \in \Gamma, \quad \exists \delta_i \in \Delta \mid \delta_i : C \times G \to \mathbb{R}_0^+$$

and

$$\forall\, \delta_i \in \Delta, \quad \exists\, d_i \in D \mid d_i : C \times G \to [0 \dots 1]$$

where $\delta_i \in \Delta$ is a distance function and $d_i \in D$ is the related inversely proportional function. For each function $d_i \in D$: $d_i(c, g)$ is equal (or tends) to 1 if $\delta_k(c, g)$ is 0 and $d_i(c, g)$

GIR '10, 18-19th Feb. 2010, Zurich, Switzerland

tends to 0 when $\delta_i(c, g)$ increases, where $c \in C$ and $g \in G$. Then, we define two auxiliary functions:

$$avg(\delta_i, c, G) = \frac{1}{\|G\|} \sum_{g \in G} \delta_i(c, g)$$

$$odf(\delta_i, c, g) = \|\{h \in G | \delta_i(c, h) \leq \delta_i(c, g)\}\|$$

The first function computes the average distance value for a given dimension, while the latter computes the number of objects with the same or shorter distance from a user for a given object. Finally, for each relevance dimension drawn from a relationship $\gamma_i \in \Gamma$, the GRBM25 model defines a similarity function as:

$$sim_i(c, g) = log \left[\frac{\|G\|}{odf(\delta_i, c, g)} \right] \cdot \cdot \frac{(k_1 + 1) \cdot d_i(c, g)}{k_1 \left((1 - b) + b \left(\frac{\delta_i(c,g)}{avg(\delta_i, c, G)} \right) \right) + d_i(c, g)}$$

where k_1 and b are tuning parameters derived from the original Okapi BM25 formula.

3. PRELIMINARY EXPERIMENTS

To achieve a first evaluation of the GRBM25 model, we tested our approach on a set of 2222 Points Of Interest (POIs) from TeleAtlas in the city of Zürich. Each POI has been modelled as a geographic object, taking into account its location (as latitude and longitude coordinates) and category (e.g. shop, restaurant, etc.). In these preliminary experiments, we modelled the user context as described by the user location (as latitude and longitude coordinates) and the user activity (e.g. shopping, touring, etc.). Finally, we establish two relationships: a spatial relationship between the user location and the object location, and a conceptual relationship between the user activity and the object category.

The distances concerning the spatial relationship have been calculated as walking distances between two locations, using the OpenStreetMap API [5]. The distances concerning the conceptual relationship have been calculated as shortest-path distances within a concept graph derived from ConceptNet [2]. Two Gaussian functions have been used as inversely proportional functions. The tuning parameters have been set to the values commonly used for the original Okapi BM25 model: $k_1 = 1.5$ and $b = 0.75$.

A first set of experiments has been carried out using a set of six different user locations and two different user activities. Correlation analysis reveal that, taking into account the whole collection, the outputs of the sim_i functions are highly correlated to the d_i values, and that there is no significant difference between the two. In other words, there is no statistical difference between the GRBM25 model and the approaches proposed in the literature.

However, considering just the most relevant geographic objects the two approaches seemed to behave differently. Thus, a second set of experiments has been carried out, taking into account four different situations, where the user was performing the same activity in four different places just slightly apart (about 25 meters) from each others. Statistical analysis of the obtained scores show that the approach based on the d_i functions notices almost no difference between the four situations. However, the scores calculated by the sim_i functions for the objects closer to the user seem to be moderately affected by those small changes, whereas the same changes do not concern at all the remaining objects.

4. CONCLUSIONS

The results of the preliminary experiments seem to imply a key difference between the proposed model and the approaches proposed in the literature [1, 4, 6]. Considering the distribution of the scores — expressed by the *avg* and *odf* functions — within the similarity computation, the GRBM25 model shows a higher sensitivity to small changes in the user context in assessing the relevance scores of the most relevant objects.

Further experiments are needed in order to better understand the behaviour of the proposed approach under different conditions and using a higher number of relevance dimensions. It is also clear that a more sensitive model has to rely on accurate information about both the user context and the geographic objects.

Future challenges will include interpretation of the user behaviour for relevance feedback. This will give us the opportunity to include relevance feedback scores in the proposed model as in the Okapi BM25 model's complete formula.

The GRBM25 is a very general model that can be applied to GIR systems, as well as to Location Based Services and Mobile Adaptive Maps. We believe that this approach could make a valuable contribution to these research fields and to the development of a reliable GR framework.

5. REFERENCES

[1] L. Andrade and M. Silva. Relevance ranking for geographic ir. *ACM GIR*, 2006.

[2] C. Havasi, R. Speer, and J. Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, 2007.

[3] K. Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval: Development and comparative experiments, Parts 1 and 2. *Information Processing and Management*, 36(6), 2000.

[4] D. Mountain. Spatial filters for mobile information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval.* ACM, 2007.

[5] OpenStreetMap. http://www.openstreetmap.org. Last accessed, December 2009.

[6] R. Purves, P. Clough, C. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. Syed, S. Vaid, et al. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7), 2007.

[7] J. Raper. Geographic relevance. *Journal of Documentation*, 63(6), 2007.

[8] T. Reichenbacher, P. Crease, and S. De Sabbata. The concept of geographic relevance. In *Proceedings of the 6th International Symposium on LBS & TeleCartography*, 2009.