Flexible Querying in Geo-Finder

Gloria Bordogna[1], Giuseppe Psaila[2]

[1] CNR-IDPA - via Pasubio 5, I-24044 Dalmine (BG) (Italy)
gloria.bordogna@idpa.cnr.it
[2] Università di Bergamo - viale Marconi 5, I-24044 Dalmine (BG) (Italy)
psaila@unibg.it

Abstract. The evaluation of queries specifying both content based conditions and spatial conditions on documents contents in Geographic Information Retrieval requires representing the vagueness and context dependency of spatial conditions and the personal user's preferences.
The *Geo-Finder* system [1] implements a Geo-Retrieval model that evaluates flexible spatial queries combined with content queries. The spatial condition is interpreted as the soft constraint "close" on the user's perceived distance. Two distinct semantics can be used to combine the spatial and the content conditions: *and possibly* or *average*; in both cases it is possible to modify the relative weight (preference) of conditions.

Keywords: Geographic Information Retrieval, Fuzzy aggregation operators, context dependent spatial query, soft constraint.

1 Introduction

An important issue in GIR is the problem of *spatial querying* [2, 5, 3], intended as *supporting the distinct information needs of users that may access the same collection for different purposes*. To address it, GIRs must be developed to take user's preferences into account, to rank query results in terms of relevance [4].

In the *Geo-Finder* system [1], we devised a Geo-Retrieval model for flexible querying a GIR, such that: the user expresses the spatial condition based on the "close" soft constraint, adapting the *spatial scope* to the perceived meaning of spatial conditions; the user expresses preferences on how to combine the content conditions with the spatial conditions.

In the spatial condition, the user's context is modeled as user's perceived distance measure, that modifies the spatial scope of the query.

Two distinct semantics are provided for flexibly combining the content condition and the spatial condition: the asymmetric *and possibly* aggregation combines the mandatory content condition with the optional spatial condition; the compensative *average* aggregation linearly combines the two conditions. The relative weight between the conditions can be specified to achieve personalization.

2 The Geo-Retrieval model

In this paper, we present the Geo-Retrieval model devised in *Geo-Finder*. It is based on the concept of *Fuzzy Footprint*, that represents the degree with which a geographic reference is relevant for a document: for each indexed document, the *Geo-Indexer* [1] generates a set of fuzzy footprints.

A *fuzzy footprint* of a document d, denoted as $Foot(d)$, is a fuzzy set of geographic coordinates $\mathtt{gc}= (lat,lon)$, where lat=latitude lon=longitude (expressed in degrees), with a membership degree $\mu_{Foot(d)}(\mathtt{gc}) \in [0,1]$ representing the significance by which the geographic location gc belongs to the geographic focus of document d:

$$Foot(d) = \{\langle\ \mathtt{gc}_1, \mu_{Foot(d)}(\mathtt{gc}_1)\rangle, \ldots, \langle \mathtt{gc}_n, \mu_{Foot(d)}(\mathtt{gc}_n)\rangle\}$$

where each $\mathtt{gc}_i = (lat_i,lon_i)$ and its membership degree $\mu_{Foot(d)}(\mathtt{gc}_i)$ are determined by the *Geo-Indexing* module [1].

A user query q consists of two conditions: a *content-based condition*, expressed by a list of content keywords, and a *spatial condition*, expressed by a list of geographic names. The spatial condition is interpreted as the requirement for documents with geographic reference "close" to the specified place names. These two conditions are evaluated by specific partial matching functions that compute two distinct scores in [0,1]: the *Retrieval Status Value* w.r.t. the content, denoted as $RSV_{content}(d)$, and the *Geographic Retrieval Value*, denoted as $GRV_{closeness}(d)$.
In *Geo-Finder*, $RSV_{content}(d)$ is a classical cosine similarity measure, computed by means of the *Lucene* library.

These two scores are finally combined to compute the *global Retrieval Status Value* w.r.t. the whole query q, indicated by $RSV_q(d)$, by applying a suitable aggregation function. We defined two aggregation functions, since we considered two distinct aggregation semantics, i.e., the *and possibly* asymmetric aggregation and the *average* compensative aggregation.

Evaluation of the spatial condition. Given the fuzzy footprint $Foot(q)$ of the geographic names in the query q, the fuzzy footprints of the documents d, $Foot(d)$, that are likely to satisfy the query are retrieved by accessing the *footprint spatial index*. The semantics of the spatial condition is that of evaluating a user's context dependent "closeness" of the documents' footprints $Foot(d)$ to the query footprint $Foot(q)$. This is done by a matching function `close` which models the concept of "close" as a user's context dependent soft constraint.

The matching function `close` computes a *Geographic Retrieval Value*, $GRV_{closeness}(d) \in [0,1]$, depending on the closeness of the document footprint to the query footprint as follows:

$$GRV_{closeness}(d) = \mu_{close}(Foot(d),Foot(q)) =$$
$$=max_{i \in Foot(d), j \in Foot(q)}\ qscope(dist(i,j) \times min(\mu_{Foot(d)}(i), \mu_{Foot(q)}(j)))$$

Where $\mu_{Foot(d)}(i)$ and $\mu_{Foot(q)}(j)$ are the membership degrees of the i-th and j-th fuzzy spatial references $\mathtt{gc}_i \in Foot(d)$ and $\mathtt{gc}_j \in Foot(q)$, i.e., the extent to which a spatial reference represents the geographic focus of the document and of the query, respectively.
The $dist(i,j)$ function is a great circle approximation of the actual distance between the two spherical coordinates $\mathtt{gc}_i$ and $\mathtt{gc}_j$.
The *qscope* function modifies the geographic distance so as to model the user perceived distance as follows:

$$qscope(x) = \begin{cases} \delta/(x+\delta) \ \ if\ x \leq \delta + k \times MaxDist(Foot(q)) & with\ \delta \geq 0, k > 0 \\ 0 \qquad\qquad otherwise \end{cases}$$

$MaxDist(X) = max_{i,j \in X}(dist(i,j))$ is the maximum geographic distance between any two geographic places i and j in the footprint X, and can be considered as the maximum dispersion of the fuzzy footprint X. It is zero in the case X contains just one single place. Thus $MaxDist(Foot(q))$ is the query dispersion. Its value depends on the number of geographic names specified in the query and on the maximum distance between their geographic coordinates.

The parameters δ and k permit to change the spatial scope of the query. The parameter δ is the *query range*, and is useful in the case of a query footprint consisting of a single geographic coordinate pair gc in order to retrieve also documents with footprint in the surrounding places. Distinct δ can adapt the evaluation of the spatial condition "close" to the user perception, thus, modeling strict or relaxed interpretations of the "closeness" surroundings of a point. The higher the δ, the greater is the surrounding.

The parameter k makes it possible to model a tolerance on the geographic distance between a document fuzzy footprint and the query footprint, so that one can consider close places within a distance of k times $MaxDist(Foot(d))$, i.e., k times the query maximum dispersion.

We consider four main query scopes that can be related to the user's context, and that are defined in the *Geo-Finder* system by the following default values of k and δ. (1) The *small scope* is defined with $k = 5$, $\delta = 3$ *km*; it is useful when $Foot(q)$ is a street address within a city or a small city and we are interested in its very near surroundings (in this case, $Foot(q)$ could vary approximately between 0 and about 10 *km*): with this setting, one can retrieve documents within a distance from the query of 3 *km* to about 50 *km*. (2) The *meso scope* is defined with $k = 4$, $\delta = 50$ *km*; in this case, $MaxDist(Foot(d))$ covers the area of either a region or a small nation like Belgium. (3) The *large scope* is defined with $k = 3$, $\delta = 1000$ *km*, in this case $MaxDist(Foot(d))$ covers the area of a medium nation such as France (in this case $Foot(q)$ could vary approximately between 0 and a few thousand kilometers). (4) The *full scope* is defined with $k = 3$, $\delta = 10000$ *km*; in thsi case, $MaxDist(Foot(d))$ covers the area of a big nation such as Russia or of a continent.

For example, if one specifies a spatial condition with the two geographic names *Bergamo*, *Como* (*Como* being at about 40 *km* from *Bergamo*), and the query scope is *meso* (i.e. $k = 4$ and $\delta = 50$ *km*) the documents with footprints at a maximum distance of 210 *km* from the query footprint are retrieved: for instance, both documents in *Milano* and *Lugano* are retrieved while a document with a footprint in *Rome* is not.

The Global RSV. *Geo-Finder* implements two distinct semantics to combine $RSV_{content}(d)$ and $GRV_{closeness}(d)$.

The asymmetric *and possibly* semantics is defined as follows:

$$RSV_q(d) = RSV_{content}(d) \textit{ and possibly}^{\alpha} \; GRV_{closeness}(d) =$$
$$= RSV_{content}(d) \times \; max((1-\alpha), GRV_{closeness}(d))$$

Parameter α specifies the user's preference of the spatial condition w.r.t. the content condition. When $\alpha = 0$, it means that the spatial condition can be disregarded to rank the documents, and in this case the *global Retrieval Status Values* is determined solely based on the content relevance score $RSV_{content}(d)$.

When $\alpha = 1$, the two conditions are both mandatory: this means that the *Geographic Retrieval Value* $GRV_{closeness}(d)$ has the same relevance of the content *Retrieval Status Value* $RSV_{content}(d)$. In this case, the aggregation reduces to the product, i.e., the "fuzzy Anding" of the two relevance scores. Intermediate values of α in $(0, 1)$ demands for an asymmetric combination. The value $(1 - \alpha)$ guarantees a minimum satisfaction level for $GRV_{closenss}(d)$, so that the spatial condition becomes optional and the global $RSV_q(d)$ is not too much penalized in the case in which the spatial condition is not satisfied.

With the symmetric *Average* semantics, the *Global RSV* is defined as follows:

$$RSV_q(d) = RSV_{content}(d)\ average^{\alpha}\ GRV_{closeness}(d) =$$
$$= (1-\alpha)\times\ RSV_{content}(d) + \alpha\times GRV_{closeness}(d)$$

When the preference degree $\alpha = 0$, the result is determined solely by the satisfaction of the content condition; conversely, when $\alpha = 1$, the global RSV is determined solely by the satisfaction of the spatial condition, and the content based condition is irrelevant. Intermediate values of α permit to vary the trade-off between the influences of the two conditions; in this case, the two conditions compensate each other, while with the *and possibly* semantics it is mandatory to satisfy the content condition to retrieve a document.

3 Conclusions

The Geo-Retrieval model described in this paper is implemented in the *Geo-Finder* system. In [1], we extensively presented its features. Furthermore, in [1], some evaluation results are also discussed showing the improvement of *Geo-Finder* ranking over *Google* ranking. The evaluations also showed that the precision of *Geo-Finder* improves when restricting the geographic domain of interest, thus outlining the positive role of modeling the user's context which determines the perceived distance when evaluating the spatial query condition.

References

1. G. Bordogna, G. Ghisalberti, and G. Psaila. Geographic information retrieval: Modeling uncertainty of user's context. *Fuzzy Sets and Systems.*
2. G. Cai. GeoVSM: An integrated retrieval model for geographic information. In *M.J. Egenhofer and D.M. Marks (Eds), GIScience 2002*, LNCS 2478, pages 65–79. 'Springer Verlag, 2002.
3. Z. Li, C. Wang, X. Xie, X. Wang, and W.Y. Ma. Indexing implicit locations for geographical information retrieval. In *n Proceedings of GIR-2006, Int. Conf. on Geographical Inf. Retrieval*, Seattle, USA, August 2006.
4. G. Mountrakis and A. Stefanidis. Moving towards personalized geospatial queries. *Journal of Geographic Information System*, 3:'334–344, 2011.
5. R.S. Purves, P. Clough, C.B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A.K. Syed, S. Vaid, and B. Yang. 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):'717–745, 2007.