Geographic Information Retrieval based on two orthogonal criteria

Gloria Bordogna [1] Giorgio Ghisalberti [2] Marco Pagani [1] Giuseppe Psaila[3]

1 CNR - IDPA, via Pasubio 5, c/o POINT, Dalmine 24044 (BG), Italy
2 Politecnico di Milano, Dip. Di Elettronica, via Bassini 16, 20133 Milano, Italy
3 Università di Bergamo, Facoltà di Ingegneria, via G. Marconi 5, 24044 Dalmine (BG) Italy
Email: gloria.bordogna@idpa.cnr.it, ghisalberti@elet.polimi.it, marco.pagani@idpa.cnr.it, psaila@unibg.it

***Abstract**— A Geographic Information Retrieval (GIR) model is defined and the architecture of a GIR system, named Geo-Finder, is described. The GIR system is constituted by two main modules. The geo-indexing module applies bipolar criteria to automatically identify the geo-reference focus (footprint) of textual documents. The geo-retrieval module applies a context dependent matching function to evaluate queries consisting of two orthogonal constraints, a content constraint and a spatial constraint; the spatial constraint is defined by modeling the user's perception of geographic "closeness" between the documents' footprint and the query footprint. For each retrieved document, two relevance scores are computed with respect to the two query conditions, that can be combined to generate an overall ranked list of documents in a flexible way, by allowing users to specify a trade off between them.*

Keywords— bipolar criteria evaluation, geo-footprint, geo-indexing, geographic information retrieval, perception distance, spatial query evaluation.

1 Introduction

The representation and management of Geographic Information is becoming a hot topic in the research area of information retrieval [9]. Given that about 15% of the queries submitted to search engines contain geographic names [20], the effective indexing and retrieval of geographical information poses new challenges to the design of *location-based search services* [18].

Location based search services allow users to search into specialized repositories, and more generally on the Internet, documents describing both resources and specific contents in the *neighbourhood of the user location*, or in the *neighbourhood of a geographic location* that is explicitly specified in the request. "*In the neighbourhood of a location*" specifies a spatial constraint on the geographic content of the retrieved documents, typically a constraint on the geographic distance from the desired location, like in the request "*find Indian restaurants near Bergamo university*".

Nevertheless, actual search engines do not enhance the influence of the geographical information content in evaluating requests containing geographic names. For example, the previous request submitted to *Google* retrieves as first ranked web page "*hotels near Indiana University East*" that does not satisfy the spatial constraint to be close to Bergamo. Generally, search engines are able to find specific resources such as hotels, restaurants, hospitals, in the neighbourhood of a locality, e.g. "*find hospital near Milan*", but this works only for specific resources, and moreover the ranking does not depend on the distance from the specified locality.

Furthermore, the indexing process does not extract most of these places automatically; consequently many web pages related to the query are not displayed as a result. In fact, many of the results obtained by a Google search are relative to resources that have directly informed Google of their geographical position. More specialized services has been developed for automatically mapping documents. For example, *MetaCarta* [11] is a system that geo-tags news, and allows searches in which one can distinctly specify the content constraint and the location of the news. However, this system is unable to retrieve news in the neighbourhood of a desired locality.

The identification of the geographic names, and successively the selection of the *geographic reference focus* (footprint) of web pages have been dealt with in several papers [1][2] [8][15]. Other papers considered the problem of spatial query evaluation [3][13][23]. In [15] the problem of implicit location identification is considered.

In this paper, after introducing the main problems involved in the design of a GIR system, we describe our GIR model and the system, named *Geo-Finder*, implementing it.

The main characteristics of the proposed geo-indexing model is the integration of multiple bipolar criteria satisfaction degrees [4][6][12][17][25], computed based on context dependent rules. Some criteria have a positive influence on the selection of the geographic names as footprints of the document; others have a negative influence. The positive and the negative constraints are heterogeneous and express pieces of information of a different nature [4]. The footprint of a document is represented as a fuzzy set of geographic coordinates (i.e., latitude and longitude, identifying the location and extent on the Earth surface of the geographic names occurring in the document) with membership degrees expressing their strength in defining the document footprint.

Second, we propose a distance measure to model the evaluation of the spatial query constraint, named "user's perception" distance. This distance measure depends on several aspects modelling the user context, such as the *spatial scope* of the query.

Finally, the system user interface maps the bi-dimensional relevance of retrieved documents in a Cartesian plane. The axes correspond to two orthogonal criteria

ISBN: 978-989-95079-6-8

(content constraint, geographic constraint), and the distance from the origin is inversely proportional to the global relevance of the documents. The user can also obtain a unique ranked list by choosing a trade-off of the two criteria satisfaction degrees based on their linear combination.

2 Why Geographic Information Retrieval is difficult

Geographic Information Retrieval (GIR) can be considered a specialized area of Information Retrieval, with an emphasis on the geographic indexing and geographic retrieval. GIR deals with any kind of information, i.e., not just maps or images but also texts, that have some relation to one or more locations on the Earth's surface, i.e., geo-referenced information [22]. Most of the information available on the Internet and in digital libraries is implicitly geo-referenced.

Often the link to the place (geographic footprint) is encoded by a geographic name. Geographic indexing implies the identification of the geographic names in a text and their translation into footprints, which are two operations that imply the management of imprecision, ambiguity, and incompleteness.

Often geographic names are ambiguous [14][21], some of them are homonymous of general terms (e.g. "*Los Angeles"*), some others identify distinct places on the Earth (e.g."*Rome*") or are temporal (e.g."*Leningrad*", "*Petersburg*", "*St Petersburg*"), or even are local names whose recognition relies on the knowledge of the local language.

Further, some pseudo-names are imprecise (e.g. "*around Milan*") or implicitly mentioned (e.g. "*the capital of Italy*"), or depending on the context (e.g. "*highest peak*" implicitly identifies "*Mont Blanc*" in a text describing the Alps). In [19] it has been argued that, given all these characteristics, geographic indexing can be feasibly faced by considering large corpora of geographic knowledge and heuristic rules.

On the other side, geographic retrieval implies being able to retrieve documents whose geographic focus satisfies a spatial constraint specified in a query. Generally, the spatial constraint demands the document footprint to be "*in the neighbourhood of a place*", i.e., it is a constraint on the distance. However, in this context the geographic distance is not merely Euclidean, but is related to the user's spatial context of interest, that can be related to the human perception of the time needed to cover it [5].

When a user searches resources or documents on the internet that are close to his/her current location, the judgment on the distance is related to his/her perception, depending on the fact that he/she is walking, driving or flying. Then, the constraint on the distance depends on the request and users' context.

Besides distance, other topological constraints could be defined, e.g., inclusion, overlapping, at the south/north of a specific region. Also these constraints must be interpreted in a tolerant, approximate way.

Then, geographic retrieval can draw benefits by defining the spatial constraints as soft, context dependent constraints, admitting degrees of satisfaction.

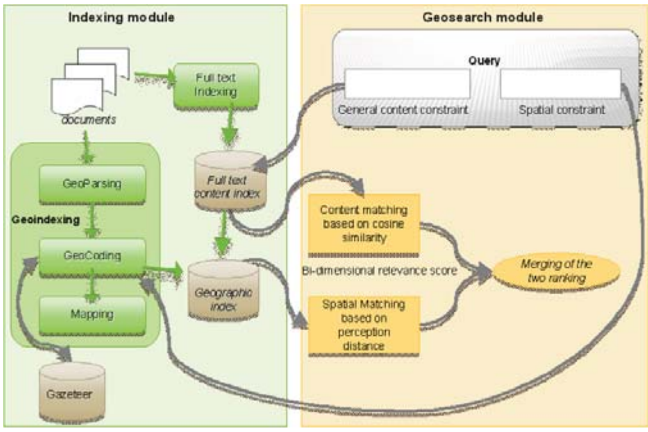

Figure1. Architecture of the Geographic Information Retrieval system *Geo-Finder*

3 The Architecture of the Geographic Information Retrieval System *Geo-Finder*

In this section we present the general architecture of the GIR system named *Geo-Finder*, that we designed and implemented. The architecture is depicted in Figure 1.

The system has the typical structure of an IRS, consisting of two main components: the *Indexing Module* and the *Retrieval Module* (named *GeoSearch*).

The Indexing Module has two main sub-modules: the *Full-Text Indexing sub-module* performs the full text indexing of the documents to represent their generic content, and generates the textual inverted index.

The *GeoIndexing sub-module* is the novel component, specialized in the identification of the footprints of documents, representing their geographic focus. This sub-module makes use of a *gazetteer* stored into a PostgreSQL database, containing the names of geographic entities (both administrative and physical entities) of all over the world, in English language and local languages (*GeoNames*) [10].

Once the *GeoIndexing* module has identified the footprints, it also stores them into the geographic index, that is a posting file containing the documents' identifiers and their footprints.

On the other side, the *GeoSearch* module interprets queries composed of two conditions. On the left hand side of the *GeoSearch* box in Figure 1, the generic content condition can be specified as a set of keywords, which define the *generic content constraint* on the documents' full text content representation. On the right hand side, the spatial condition can be specified as a geographic name, which is used to define the *spatial constraint* on the documents' footprints.

The *GeoIndexing* module parses the terms in the spatial condition to identify their footprint. These two conditions are evaluated by the sub-modules named *Content Matching module* (based on the *Lucene* library) and *Spatial Matching module*, based on perception distance (that is based on the original geo-retrieval model introduced in the next section), respectively.

In order to be retrieved, a document must satisfy, at least a little, the first content based condition (necessary condition), while the spatial condition is used for conditioning the ranking of the documents (optional

ISBN: 978-989-95079-6-8

condition). Then, the two conditions are *merged* by an "*and possibly*" aggregator, applied by the *Merging* sub-module that combines the satisfaction degrees of the two constraints, in order to obtain a unique value that performs the global ranking of documents. In the combination, one can give more preference to one of the two conditions, in order to emphasize the influence on the global ranking of either the document content or the document footprint.

4 The Geo-Retrieval Model

In this section, we describe the geo-retrieval model at the basis of the system implementation. In the first subsection the *GeoIndexing* model is described. In the second subsection, we introduce the geo-retrieval model at the basis of the *GeoSearch* component.

4.1 The GeoIndexing model

The GeoIndexing model is defined to identify for each document its footprint. A footprint of a document *d*, *Foot(d)*, is as a fuzzy set of geographic coordinates gc=(*lat,lon*), *lat*=latitude *lon*=longitude, expressed in degrees, with a membership degree $\mu_{Foot(d)}(gc)=GeoRef(gc)\in[0,1]$ representing the strength by which the geographic location gc, named gw, belongs to the footprint of the document d.

$$Foot(d)=\{GeoRef(gc_1)/gc_1,\ldots,GeoRef(gc_n)/gc_n\}$$

A document is represented as a stream of tokens <*t*>. Some terms *t* have been selected as content indexes and thus are in the dictionary. For each of them, the frequency in the collection is known, while in the posting list we enter the documents in which they appear with their significance degrees *F(d,t).* The significance degree is usually defined based on statistic analysis of the document text [16].

Besides this information, in the posting list we can also find the positions of the occurrences of the index terms in the document text, $occ_k(t_i, d)$, the k-th occurrence of t_i in d.

The identification of the document footprint is achieved in three steps, where distinct sets of heuristic rules are evaluated. Each set of rules acts as a filter on the input terms, so that only those terms whose global satisfaction degree of the set of rules is above a threshold are selected as input to the second step.

The *GeoIndexing* module implements this geoindexing model. It operates in two subsequent phases: first, it performs a G*eoParsing* that applies the first two sets of rules to detect the candidate geographic names (gw). Then, the *GeoCoding* sub-module identifies the document footprint *Foot*(d) and stores it in a file with the document unique identifier d.

The first set of rules consists of Name Entity Recognition (NER) rules, aimed at reducing the set of terms among which to successively select the candidate geographic names e.g.:

```
if Language(d)="English"∧
   FirstChar(t)=Capital then return(t).
```

The second set of rules receives, in input, a stream of previously selected terms, hereafter indicated by gw, and filters the candidate geographic names, a subset of the input terms. It applies bipolar [4] context dependent rules (r_i) exploiting a *gazetteer* [10] and computing independently a satisfaction degree *s(*gw*)* and a dissatisfaction degree *d(*gw*)* that denote to what extent gw is a geo-name and is not a geo-name, respectively.

The aggregation of the positive (negative) rules is done based on a Generalized Conjunction Disjunction function (GCD) [12], that, for distinct values of the parameter *p*, can model aggregations from completely compensative (or), where each rule can replace any other, to completely not compensative (and), where all rule must be satisfied simultaneously [6]:

$$s(gw)=\left(\sum_{i=1}^{m}\lambda_i*\left(r_i\,(gw)\right)^{ps}\right)^{1/ps} \qquad d(gw)=\left(\sum_{j=1}^{n}\overline{\lambda}_j*\left(\overline{r}_j\,(gw)\right)^{pd}\right)^{1/pd} \tag{1}$$

The rules r_i and $\overline{r}_j$ assume a value in [0,1]. *ps* and *pd* are set so as to define (partially) compensative aggregations. In our experiment we used *ps*=*pd*=20, i.e., towards or like aggregation [6].

$\lambda_i,\overline{\lambda}_j\in[0,1]$ with $\sum_{i=1}^{m}\lambda_i=1,\sum_{i=1}^{n}\overline{\lambda}_j=1$ are the weights of the rules, i.e. their importance degrees in the aggregation, and are determined based on statistical analysis on a sample set of documents of a collection, and are set in a configuration file that is read by the *GeoParsing* sub-module during index generation. This way, the geoindexing can be suited to the characteristics of a collection.

The satisfaction of a rule r_i with *i=1,m* is interpreted as a hint of evidence that gw is a geographic name; thus the first *m* rules have a positive influence on the recognition of gw as a candidate geoname, like, e.g., the following two rules:

```
If gw ∈ gazetteer ∧ ∃gwk ∈ d
∧ gwk=administrative_distr(gw)⇒r2(gw)=0.5
∧|occi(gw,d)-occj(gwk,d)|<Δ⇒r2(gw)=r2(gw)+0.5
```

E.g.if gw="*San Francisco*"and gw_k="*California*", its administrative district, occurs in the same document d at a maximum distance *Δ=3* words, then r_2(gw)=1, otherwise if the occurrences are at a greater distance than *Δ* the rule is only partially satisfied, i.e., r_2(gw)=0.5.

Another rule is the following:

```
If gw ∈ gazetteer ∧ gw-1 ∈ prefix ⇒ r3(gw)=1
```

e.g. if gw="*Blanc*" is preceded by gw_{-1}="*Mount*" that belongs to the set prefix of prefixes of geographic names such as *Mount*, *lake*, *city*, *river*, then r_3(gw)=1.

Conversely, the satisfaction of a rule $\overline{r}_j$ with *j=1,n* is interpreted as a hint of evidence that gw is not a geographic name. Thus, these rules have a negative influence, like the following one:

```
If gw ∈ gazetteer ∧ gw ∈ Stopwords ⇒r̄4(gw)=1
```

e.g. `gw ="Nice"` is a `stop-word` too.

For each input `gw`, a geo-score *GeoScore*(`gw`)∈[0,1] is computed based on the values of *s*(`gw`) and *d*(`gw`) as follows:

$$GeoScore(\mathtt{gw}) = \begin{cases} s(\mathtt{gw}) - d(\mathtt{gw}) & if\ \ s(\mathtt{gw}) > d(\mathtt{gw}) \\ 0 & if\ \ otherwise \end{cases}$$

We select the `gw` with *GeoScore*(`gw`)$>\tau>0$ as reliable geographic names. This threshold allows restricting the footprint of a document to reduce the possibility of identifying false positives. τ must be set based on experimentations, and is also specified in the configuration file. The greater it is, the smaller is the possibility of selecting false geo-names. Generally, it is better to lose some true geo-names than to select false ones.

The third set of rules, used in the third step by the *GeoCoding* sub-module, is aimed at identifying, from the reliable geographic names, the geographic locations that belong to the document footprint (i.e., the fuzzy set *Foot(d)* of pairs of geographic coordinates).

For each of the selected geo-names, `gw`, a (set) of pair(s) of geographic coordinates $GC_{gw}=\{gc_1,\ldots,gc_n\}$, with `gc`=(`lat,lon`), is retrieved from the `gazetteer`. Each pair `gc` is a geocode of `gw`, uniquely identifying a geographic place on the geographic domain. Notice that homonymous geo-names have the same `name` but distinct `gc` pairs.

For each geocode associated with a selected `gw`, a *geo-reference score (GeoRef*(`gc`)*)*∈[0,1] is computed, that expresses the strength by which `gw`, located in `gc`, belongs to the document footprint (i.e. it is marginal or central in defining the geographic focus of the document).

A true geographic name can be correctly identified in a document, but it can be meaningless in defining the geographic focus of the document itself. Let us consider, for example, the geographic names that are often present at the very end of web pages or in their footnotes: they generally have nothing to do with the document content, but are related with the affiliation of the web master, and thus must not define the document footprint.

Also these rules have a satisfaction degree $r'_i(gc)\in$ [0,1] that is dependent on some variable. For example, the frequency $F(d,gw)$ of a geoname `gw` in a document *d*, increases the strength of its geocodes in the footprint proportionally to the degree by which the geoname was recognized as a candidate geo-name *s*(`gw`), and inversely proportional to the degree by which it was recognized as not being a geo-name *d*(`gw`). This is represented by the following rule:

$$r'_1(gc) = \mu_{significant}(F(d,gw) * GeoScore(gw))$$

with $gc\in GC_{gw}$ and $\mu_{significant}$ is a monotonic non-decreasing membership function.

Another rule evaluates the presence of another geoname gw_k in *d*, whose geo-code is geographically near to that of `gw` (at a maximum geographic distance `dist` equal top *Δ*). When this occurs, it increases *GeoRef*(`gc`).That is:

$$r'_2(gc) = max(0, \Delta - dist(gc, gc_k))$$

with $gc\in GC_{gw}$, $gc_k\in GC_{gwk}$ and $gw_k \in d$.

The population attribute of an administrative name `gw`, that has multiple geocodes, indicated by *population*(`gc`), is used to favour the influence of big cities w.r.t. small ones having the same geoname in the footprint:

$$r'_3(gc) = (population(gc))/max_k(population(gc_k))$$
$$gw \in gazetteer,\ gc, gc_k \in GC_{gw}$$

This rule allows us to resolve ambiguities, generated by homonymous geographic names, that is, to favour the geographic coordinates (geo-codes) that most likely belong to the document footprint.

These rules are aggregated based on a GCD aggregation, in which each rule has a weight $\lambda'_i\in$[0,1], with $\sum_{i=1}^{m}\lambda'_i = 1$, determined based on statistical analysis and the parameter *p*=1, neutral aggregation (all configurable):

$$GeoRef(gc) = \left(\sum_{i=1}^{m}\lambda'_i\left(r'_i(gc)\right)^p\right)^{1/p} \qquad (2)$$

A minimum threshold *Φ*>0 on *GeoRef*(`gc`) (configurable) restricts the footprint of a document to reduce its extent. Therefore, at the end of this model, for each geoname we have two different scores: *GeoScore*(`gw`)$>\tau>0$ and *GeoRef*(`gc`)$>\Phi>0$. Figure 2 shows the geonames extracted from a sample paper. The map was generated by the *Mapping module*: this module creates distinct graphic metafiles for each document to map the footprint in distinct graphic environments, such as Gmaps for Google maps (see figure 2), KML for Google Earth, and GPX for GPS data format environments .

Score	GeoLocation	Name	Latitude	Longitude	Country code	Population
0,54	Milano	Milan	45.461216	9.187832	IT	1306661
1,00	Italy	Italian Republic	42.833332	12.833333	IT	58103033
0,87	Abano	Abano Terme	45.360992	11.791407	IT	18206
0,76	Ferrara	Ferrara	44.82678	11.620713	IT	130992
0,78	Vicenza	Vicenza	45.557293	11.540898	IT	107223
0,92	Tuscany	Tuscany	43.416668	11.0	IT	3573454
0,63	Apulia	Andria	41.23	16.29	IT	4021957

Figure 2. Google map of a document footprint, generated by *Geo-Finder* mapping module. The hue of the red pins represents the degree of the *geo-Reference* scores of the localities in the document footprint. A table with the geographic coordinates in the footprint and their country code and population attribute is also reported.

ISBN: 978-989-95079-6-8

4.2 The GeoRetrieval model

The GeoRetrieval model takes the footprint (*Foot(q)*) of the spatial query condition *q*, that consists of a single or a set of geographic names. The footprint *Foot(q),* if not directly found in the geographic index, is identified by applying the *GeoParsing* and *GeoCoding* rules described in the previous subsection.

For each document d that has been retrieved by the *Content-based Matching* module, we match its footprint*, Foot(d)* w.r.t. the query Footprint *Foot(q)* by applying formula (3). This matching function defines the semantics of the soft spatial constraint "*close*" that computes a degree of satisfaction *GRS(d)*$\in[0,1]$ (*GeoRelevanceScore*) as follows:

$$GRS(d) = \mu_{close}\left(Foot(d), Foot(q)\right) \tag{3}$$

$$\mu_{close}(Foot(d), Foot(q)) =$$

$$= \max_{i,j}\left[\mu_{Foot(d)}[i] * \mu_{Foot(q)}[j] * \left(1 - \frac{qscope(dist(i,j))}{Maxdist(Foot(d) \cup Food(q))}\right)\right]$$

with

$$\mu_{Foot(e)}[i] = GeoRef(gc_i) > 0$$

and

$$qscope(x) = \begin{cases} x & if \quad x \le \delta + k * Maxdist\ (Foot(q)) \\ & \quad with \qquad \delta \ge 0 \;\; k > 0 \\ 0 & \quad otherwise \end{cases}$$

and *Maxdist(X)* is the maximum geographic distance (*dist*) of any two elements in the footprint *X*.

i and *j* represent the i-th and j-th pairs of geographic coordinates *latitude* and *longitude* in the footprints of the document *d,* with membership degrees *GeoRef*(*i*), and of the query *q*, with membership degrees *GeoRef*(*j*) that is assumed equal to 1.

The *qscope* function models the "user perception" distance measure, where *Maxdist(Foot(q))* is the query footprint maximum dispersion. δ and *k* are constant values that define the query scope and are set without user explicit input.

We consider four main query scopes, a *full* scope considering the whole globe, a *large* scope, considering an area covered by a continent or a big country like Russia, a *meso* scope considering an area of a nation or a big region, and a *small* scope considering a city and its surroundings. Each scope has specific values for the parameters (e.g. *k=3,* δ*=10km* is associated with a small scope, *k=30*, δ*=100km* with a *meso* scope).

δ is the query *range*, and is useful in the case of a query footprint consisting of a single geographic coordinate pair in order to retrieve also documents with footprint in the surrounding places. Distinct δ can adapt the evaluation of the spatial constraint "*close*" to the user perception; thus, modelling strict or relaxed interpretations of the *close* surrounding of a point. *k* allows to model a tolerance on the geographic distance between a document footprint and the query footprint that is equal to *k* times the query maximum dispersion, i.e., *Maxdist(Foot(q))*. This allows enlarging or reducing the query scope. This parameter can be related to the scale of the map needed to represent the minimum bounding box of the Minkowski sum of *Foot(q)* and a circle of radius *k* [7].

For example, if one specifies the two geonames *Bergamo*, *Como* (*Como* being at about 40km from Bergamo*)* as spatial condition, and the query scope is *small* (i.e. *k=3* and δ*=10km*) documents with footprints at a maximum distance of 130 km from the query footprint are retrieved (e.g. both documents in *Milano* and *Lugano* are retrieved while a document with a footprint in *Rome* is not).

On the other side, a query with footprint in *Bergamo, Dalmine* (10 *km* from *Bergamo* in *Milano* direction) will retrieve documents at a maximum distance from the query footprint of 40 km, (e.g. it will retrieve just the document in "*Milano*" and not the one in "*Lugano*").

Figure 3 depicts the *GeoSearch* user interface of the GIR system. At the top, there are two text forms for submitting the content (left) and spatial (right) query condition.

Bottom, on the right panel the bi-dimensional relevance domain is depicted, in with each point corresponds to a retrieved document. The origin identifies the query. The document's X coordinate (Y coordinate) is its relevance degree w.r.t. spatial (general content) query constraint. The closest the document is to the origin, the most relevant it is with respect to at least one query constraint.

On the left panel, the ordered list of retrieved documents is reported corresponding with a merging of the two relevance degrees giving equal importance to the two conditions. By moving the sliding bar at the top of this panel, it is possible to modify the preference between the two conditions and thus to re-rank the documents accordingly. This is achieved by a linear combination of the relevance rankings.

5 Conclusions

The system has undergone a first evaluation based on a collection of 1100 documents in *Italian* and *English* with an average length of 800 words. The collection comprehends research papers on geological studies carried out at IDPA CNR, Reuters news of the RCV1 collection, and web pages of the Open Directory Project.

This first evaluation was aimed at estimating the ability of the *GeoParser* to identify the correct geonames in documents text. For each document, we classified its geo-names and we compared the classification with respect to the footprints of the documents. We achieved a recall of 91% and a precision of 93%. Nevertheless, these results are preliminary, and further tests are needed to complete the evaluation.

The contributions of this proposal with respect to current practice are several: first of all, the computation of a fuzzy footprint to represent the geo-reference focus of a textual document, based on a bipolar criteria decision process; second, the use of a user's perception based distance measure at the basis of the computation of the degree of satisfaction of the spatial query constraint; third, the evaluation of a bi-dimensional relevance score, and the possibility to flexibly merge the two rankings into a single one by specifying a relative importance weight of the two query constraints.

ISBN: 978-989-95079-6-8

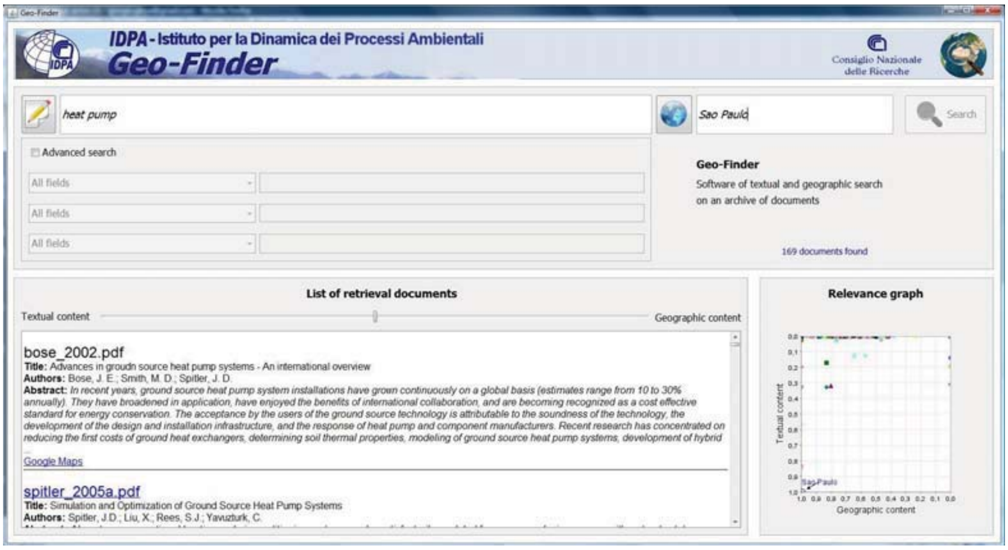

Figure 3. *GeoSearch* user interface of *Geo-Finder*: at the top the two query fields for specifying the content based condition="*heat pump*" and spatial condition="*Sao Paulo*"); below the two results panels. On the right side the bi-dimensional relevance graph, on the left the ranked list (merging with equal priority the two relevance scores).

References

[1] Amitay E., Har N., et al Web-a-where: Geotagging Web Content. SIGIR 2004.

[2] C. Wang, X. Xie et al., Detecting Geographical Locations from Web Resources. GIR'05, Bremen, Germany, 2005.

[3] Cai, G., "GeoVSM: An integrated Retrieval model for geographic information" M.J. Egenhofer and D.M. Marks Eds: GIScience 2002, LNCS 2478, Springer Verlag, 65-79, 2002.

[4] D. Dubois, H. Prade, An Introduction to Bipolar Representations of Information and Preference, Int. J. of Intelligent Systems, 23, 866–877, 2008.

[5] Egenhofer M., J., Mark D., M., Naïve Geography, eds Frank A.U., Kuhn W., Spatial Information Theory: a theoretical foundation for GIS . Springer Verlag LNCS 998, 1-15. 1995,

[6] G.Bordogna, G. Psaila, Modeling Soft Conditions with Unequal Importance in Fuzzy Databases based on the Vector p-norm, in Proc. of the IPMU Malaga, 2008.

[7] G.Bordogna, M. Pagani, G.Pasi, G. Psaila, Modeling uncertainty in Location-based spatial queries evaluation, Fuzzy Sets and Systems, accepted, september 2008.

[8] Hill L., "Introduction to Georeferencing in Digital Libraries Tutorial Document", University of California, Santa Barbara Aug. 2003.

[9] Himmelstein M.. Local search: The internet is the yellow pages. Computer, 38(2):26–34, 2005.

[10] http://www.library.uu.nl/kaartenzaal/Softwaresite/Beschrijvingen/geoname.html.

[11] http://www.metacarta.com/

[12] J.J. Dujmovic´, H. Larsen, Properties and modelingof partial conjunction/disjunction, in: B. De Baets (Ed.), Proceedings of Eurofuse Workshop on Data and Knowledge Engineering. Published as Current Issues in Data and Knowledge Engineering, Exit Publishers, Warsaw, 215–224, 2004.

[13] Jones et al, . "Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project" SIGIR 2002: Proc. of the 25th ACM SIGIR Conf.,Tampere, Finland', ACM Press, 387 – 388, 2002.

[14] Li H., Srihari R.K., Niu C., Li W., Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In Proc. of the HLT-NAACL 2003 Workshop on Analysis of geographic references, 39-44.2003.

[15] Li Z., Wang C., Xie X., Wang X., Ma W.Y., Indexing implicit locations for geographical information retrieval, in Proc. of GIR'06, August 10–11, 2006, Seattle, USA, 2006.

[16] Lucene library http://lucene.apache.org/java/docs/

[17] Matthé T., De Tré G., (2009) Bipolar Query Satisfaction Using Satisfaction and Dissatisfaction Degrees, Proc. of the ACM SAC 2009, IAR track, 1699-1703.

[18] Mountain D. and MacFarlane A.. Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals. Journal of Information Science, 33(5):515–530, 2007.

[19] Reid J., A., "geoXwalk – A Gazetteer Server and Service for UK Academia" Koch/Sølvberg (Eds) Research and Advanced Technology for Digital Libraries: in Proc. of the 7th European Conf. ECDL 2003 Trondheim, Norway, August 17-22, 2003.

[20] Sanderson M. and Kohler J.. Analyzing geographic queries. In Workshop on Geographic Information Retrieval SIGIR, 2004.

[21] Smith D.A., Crane G., Disambiguating geographic names in historical digital library. In P. Costantoupolos and I. Solvberg eds, ECDL, LNCS 2163, Springer Verlag, 127-136, 2001.

[22] Vestavik Ø., Geographic Information Retrieval: an Overview, 2008.www.idi.ntnu.no/~oyvindve/article.pdf,

[23] Woodruff A.G., Plaunt C., GIPSY: Georeferenced Information Processing System, Journal of the American Society for Information Science, 1-21, 1994.

[24] Yasuda N., Toda H., Hirano T., Geographic Information Retrieval for just your surroundings, in Proc of the WWW08, April 21-25, Beijing China, 2008

[25] Zadrozny S., Kacprzyk J., (2007) Bipolar Queries Using Various Interpretations of Logical Connectives. IFSA (1), 181-190.

ISBN: 978-989-95079-6-8