Geographic information retrieval: Modeling uncertainty of user’s context

Gloria Bordogna\textsuperscript{a,}*, Giorgio Ghisalberti\textsuperscript{b}, Giuseppe Psaila\textsuperscript{c}

\textsuperscript{a} CNR - IDPA, via Pasubio 5, c/o POINT, Dalmine 24044 (BG), Italy
\textsuperscript{b} Politecnico di Milano, piazzale Leonardo da Vinci, 20133 Milano, Italy
\textsuperscript{c} Università di Bergamo, Facoltà di Ingegneria, via G. Marconi 5, 24044 Dalmine (BG), Italy

Available online 22 April 2011

Abstract

Geographic information retrieval (GIR) is nowadays a hot research issue that involves the management of uncertainty and imprecision and the modeling of user preferences and context. Indexing the geographic content of documents implies dealing with the ambiguity, synonymy and homonymy of geographic names in texts. On the other side, the evaluation of queries specifying both content based conditions and spatial conditions on documents’ contents requires representing the vagueness and context dependency of spatial conditions and the personal user’s preferences. The spatial condition can be specified linguistically in the query through vague terms such as “close to the North East of Milan”, whose semantic depends on the user’s context and perception of distance. Further, users may want to express queries in which the content condition and the spatial condition have a distinct preference and are combined with a distinct semantics. In this paper, we propose a geographic information retrieval model and a system implementing it that represents both the uncertainty in indexing the geographic documents’ content and the user’s context and preferences in evaluating flexible spatial queries. It extracts the geographic content from documents’ text by applying heuristic knowledge coded by bipolar rules which evaluate positive hints and negative hints for the recognition of geographic names in text. Thus, it represents the geographic content of documents by fuzzy footprints, i.e., distinct locations on the earth associated with the text with a distinct degree of significance. Finally, the system allows evaluating two types of queries flexibly combining the content based condition with the spatial condition. The spatial condition is interpreted as the soft constraint “close” on the user’s perceived distance between the documents’ footprint and query’s footprint. For each retrieved document, two relevance scores are computed with respect to the two query conditions that are flexibly combined to generate an overall ranked list of documents. The user can choose the semantic for the combination that can be either an asymmetric “and possibly” aggregation between the mandatory content condition and the optional spatial condition, or a compensative “average” aggregation, defined as a linear combination of the two conditions; further, a relative preference between the conditions can be specified to achieve personalization and effectiveness. A prototypal geographic information retrieval system, named Geo-Finder, based on this model is described, and its evaluations are discussed.

© 2011 Elsevier B.V. All rights reserved.

Keywords: Geographic information retrieval; Fuzzy aggregation operators; Bipolar criteria evaluation; Geographic footprint; Context dependent spatial query; Soft constraint

* Corresponding author.
E-mail address: gloria.bordogna@idpa.cnr.it (G. Bordogna).

0165-0114/ - see front matter © 2011 Elsevier B.V. All rights reserved.
1. Introduction

Geographic information retrieval (GIR) is a specialization of information retrieval (IR) for it is concerned with indexing, searching, retrieving and browsing of geo-referenced information, with an emphasis on spatially and geographically focused indexing and retrieval.

Accessing information by its geographic reference can be natural and useful in several contexts: for example when looking for resources on a territory for tourist purposes and from mobile devices it is common practice to access relevant information by specifying the geographic place of interest; when planning rescue operations during emergencies caused by natural disasters is of great usefulness retrieving volunteered geographic information created freely by testimonies of the events [16]. More generally, it has been estimated that about 15% of the queries submitted to general purpose search engines contain geographic names [33]. Thus the representation and retrieval of geographic information is nowadays a hot research topic [14].

As defined in [19], the research on GIR combines aspects of the research on IR, DBMS, user interfaces, GIS, and the design of GIR systems that are effective and efficient.

Current approaches to GIR are diversified, ranging from basic IR approaches with no concern for spatial and geographic indexing and reasoning, like generic search engines do, to more specialized approaches applying NLP part-of-speech tagging and geographic name entity recognition for extracting geographic names and spatial relationships from the texts and queries [12,29,34]. More recent proposals apply machine learning and exploit knowledge bases such as geographic thesauri and gazetteers for the resolution of geographic ambiguity, for gazetteer construction (GNIS, World Gazetteer) and for query expansion [1,18,24,31,32,40]. Texts geo-tagging, and successively the selection of the geographic reference focus (footprint) of web pages have been dealt with in several works [1,13,21,36]. Other works considered the problem of spatial query evaluation [7,18,37] and identification of implicit locations in texts [21].

The SPIRIT (spatially aware information retrieval on the Internet) European project mainly focused the improvement of GIR on the Internet by the aid of geographic ontologies, considered as an extension of geographic gazetteers, to expand the vocabulary of place-names and to conceptualize the space structure.

GIPSY [37] is a geo-referenced information processing system relying on gazetteers to identify geographic names in texts and to associate their footprints, i.e., their geographic coordinates on Earth. MetaCarta [26] is a GIR system that geo-tags news, and permits searches in which one can specify the content condition, the desired geographic location of the news, and the temporal interval of relevance. This system also allows personalizing the results’ ranking: one can choose to order the news retrieved either in ascending or descending order of the satisfaction degree of one of the three conditions, i.e., content relevance, distance form desired location, and time of the news. Nevertheless, it does not model the user’s context, nor it permits ranking of the news based on a trade off of the distinct criteria.

In the recent past the evaluation forum of the GeoCLEF 2008 [23] highlighted many open issues relative to GIR research, among which the association of a meaning to both vague spatial conditions, i.e., spatial conditions that admit degrees of satisfaction [5] such as “close”, and vague spatial locations, i.e., areas with broad boundaries [3] such as “Mid Europe”. In fact, as pointed out in [35], one main challenge of GIR is modelling the human cognitive perception of spatial information: the way humans learn geographic knowledge, named naïve geography [11], is both by a procedural approach, for example travelling around, and by looking at maps. Taking into account the main characteristics of such knowledge is important to design more effective GIR systems that pose new challenges [27]: GIR systems should allow users to search the Internet for documents on both resources and specific contents in the neighbourhood of a desired location that is explicitly specified in the request or is detected by the GPS device of the user’s cellular phone. “In the neighbourhood” is just an example of a context dependent spatial condition on the distance from a location that one can submit to a GIR system, like in the request “find Indian restaurants close to Bergamo University”. Other spatial conditions might be topological and directional ones such as “within a region”, “North_West of a location”, etc. One characteristic of such spatial conditions is that people perceive distance in relation to the effort over time spent in travelling from a point to another. Clearly moving by plane or on foot makes a great difference on the perceived notion of closeness between the two points.

In this paper, we propose a flexible GIR model that, on one side, allows representing the uncertainty of the geographic content of texts through the association of fuzzy footprints, i.e., distinct locations on the earth associated with the text with a distinct degree of significance, and, on the other side, it allows both managing the user’s perceived meaning of spatial conditions expressed in a query through soft constraints and the user’s preferences when combining the content conditions with the spatial condition in queries. To our knowledge there is no GIR system that both represents
the uncertainty of the geographic footprints of textual documents and models the user’s perceived meaning of spatial query conditions within flexible queries.

The main characteristic of the proposed geographic indexing model is the evaluation of multiple bipolar criteria \([5,9,10,25,38]\) 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.

Second, a user’s perceived distance measure that modifies the spatial scope of the query is proposed in order to model context dependent spatial query conditions.

Third, two distinct semantics are allowed for combining the content condition and the spatial condition expressed in a flexible queries: the user can choose both an asymmetric “and possibly” aggregation between the mandatory content condition and the optional spatial condition, or a compensative “average” aggregation defined as a linear combination of the two conditions. Further the relative preference between the conditions can be specified so as to achieve personalization.

The paper is organized as follows: Section 2 introduces the main problems of geographic information retrieval and reviews the literature related to our proposal. Section 3 illustrates the general architecture and characteristic of the Geo-Finder system, implementing the proposed model. Section 4 briefly introduces the formal framework of generalized conjunction and disjunction functions used to model the rules aggregation. Section 5 is the core of the paper describing the geo-indexing model (Section 5.1) and the Geo-retrieval model (Section 5.2). Finally, Section 6, discusses some evaluations of the proposed model performed with the use of Geo-Finder system. The conclusions summarize the main achievements.

2. Main problems of geographic information 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 \([35]\). Most of the information available on the Internet and in digital libraries is implicitly geo-referenced.

As stated in \([17]\) there are still several aspects of GIR that constitute significant research challenges among which: detecting geographic content within text documents and in users queries by disambiguating their geographic location; geometric context dependent interpretation of the meaning of vague spatial conditions such as “close”; ranking the relevance of documents with respect to their geographic content as well as their thematic content.

The geographic content of a web page may have distinct meaning \([1]\): it may refer to the physical location of the server storing the web page or the address of its author or owner, or it may refer to the contents of the page and relates to the topic the page is discussing.

Automatically detecting the geographic content of textual documents, i.e., geographic indexing, often means identifying the geographic names in the document text that encode the link to a place on Earth (geographic footprint).

Even in cases where a document is accurately manually indexed, geographic names, consisting of text strings, have several well-documented problems with ambiguity, synonymy, and with name changes over time. Often, geographic names are ambiguous: they may refer to personal surnames in the text \([20,34]\) and they can be homonymous of general terms (e.g., “Los Angeles”). Many geographic names identify distinct places on the Earth (e.g., “Rome”, “Paris”), some others may have changed over time (e.g., “Leningrad”, “Petersburg”, “St Petersburg”); many places have several names in distinct languages 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 \([34]\) it was estimated that 92% of names in their corpus is ambiguous, and \([1]\) reported that 37% of the potential geographic name mentions in web pages have several possible geographic meanings.

Despite the fact that geographic indexing is not easy it is an essential task of a GIR system. In fact, encoding the geographic content of a document by geographic coordinates has several advantages over geographic names: the coordinates are persistent regardless of the changing of the name. They can be simply encoded in a format suitable for spatial browsing interfaces used by GISs so that they can be easily displayed, zoomed and panned on digital maps. Finally, they provide a consistent framework for evaluating spatial queries: having geographic coordinates of a place as index entries (either a point or a polygon) allows both precise and approximate evaluation of spatial conditions based on distance, topological, and directional selection conditions, by exploiting spatial query operators which are typical of GISs.
To face geographic indexing in the literature three main approaches have been followed: the first one uses NLP techniques [8,12,29,34], the second applies machine learning and data mining [31,40], and the third one exploits geographic ontologies and gazetteers [1,24,18,32,39].

Our approach is hybrid for it exploits both name entity recognition, statistical text analysis, and a geographic knowledge base consisting of a gazetteer, several lists of geographic mark up terms helping to identify and disambiguate geo/non-geo terms, and contextual heuristic rules coding geographic reasoning. The main characteristic of the proposed approach is that the geographic footprint of a document is identified by estimating for each of its locations a significance degree in [0,1], hence we name it fuzzy footprint, that reflects the plausibility of the identification of the location as meaningful geographic content of the document. The results of the geo-indexing can be tuned by specifying minimum desired levels for the significance degrees, thus obtaining less or more focused geographic footprints.

Furthermore, geographic retrieval implies being able to model the user’s context [35]. In fact, when retrieving documents by evaluating a spatial query like “in the neighbourhood of a place”, it is necessary to define the meaning of the vague spatial condition “in the neighbourhood”, that is defined on the distance between geographic places. However, this distance is not merely Euclidean, but the user perceives it in a different way depending on several factors related to the type of the specified place (city, province, nations, etc.) and his/her spatial context of interest: for example, it can be related to the time needed to cover it [11] which depends on the fact that he/she is walking, driving or flying.

Specifically, by taking ideas from a previous work [4] our proposal adapts the interpretation of the spatial condition “closeness” so as to model a user’s perceived distance.

Further the satisfaction of a global query submitted to a GIR system does not depend merely on the satisfaction of the spatial condition itself, but it depends on the satisfaction of a content condition as well. In fact, generally one does not want to retrieve any document related to a location, but documents dealing with interesting contents. For example, a user looking for historical interesting places close to a locality, may want to order the results based on a trade off of satisfaction of the two conditions (not too far away but interesting too). For example, a user very interested in Roman archaeological sites may tolerate a little more distance to visit a very interesting Roman settlement than a non-culturally motivated user who, instead, might prefer minimizing the driving distance for planning his/her tour.

Our approach performs a flexible combination of the content condition and the spatial condition in a query so as to allow users to express a preference between the two conditions based on their personal interests.

Current GIR systems do not model the contextual dependency of spatial conditions in queries, neither the flexible aggregation of the content conditions and spatial conditions. The GIR system Metacarta [26] only allows ranking the news based on one single criterion (either content relevance, distance for a desired location, timeliness). The spatial condition is not contextualized with respect to the user, but depends on the type of the desired location, so that continents, nations and cities are associated with a distinct range query.

Also the GIR systems that participated to the evaluation forum of GeoCLEF 2008 [23] do not model the user’s context dependency of vague spatial conditions, and do not provide customizable flexible aggregations of content and spatial relevance degrees, but apply a fixed aggregation function.

In this paper we focus on both a flexible indexing of the documents’ geographic content, and the modelling of the context-dependent meaning of the spatial query condition and its soft aggregation with the content condition. The spatial condition is defined as a soft constraint depending on a user’s perceived distance, and its combination with the content condition is modelled by two aggregation operators with preferences: an asymmetric aggregation of a mandatory and a desired condition, and a symmetric compensative weighted aggregation.

3. The architecture of Geo-Finder, a geographic information retrieval system managing uncertain information

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 Fig. 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 geo-indexing sub-module is the novel component specialized in the identification of the fuzzy footprints of documents, representing their geographic focus. This sub-module makes use of a geographic knowledge base, stored
into a PostgreSQL database extended with PostGIS 1.3 to manage the spatial components [2], and a heuristic knowledge base consisting of a library of functions. They contain distinct kinds of information:

- a local gazetteer in English and local languages that contains both names of administrative and physical geographic entities of all over the world with distinct level of granularity (cities, regions, states, nations, rivers, mountains, lakes, street addresses, etc.) and points of interest (such as the statue of Liberty in NYC, the Eiffel Tour in Paris, the Colosseum in Rome, etc.), their geographic coordinates and a set of thematic information on the type of the entity, other specific attributes such as population, and their topological inclusion relations with other entities (for example, a city is reported with its province, region, state and nation). The local gazetteer is periodically updated with data from free online sources such as GeoNames [15] and data from Opens Street Map [28].

- Distinct lists of geographic mark up terms (such as city, mountain, street, etc.), categorized for administrative, physical and POIs, that often appear adjacent in the text to the geographic names.

- Three distinct sets of parameterised rules, constituting the library of functions that encode the heuristic knowledge necessary to the geo-indexing sub-module to cope with both the geo/non-geo ambiguities and geo/geo ambiguities of geographic information in a text. Geo/non-geo ambiguity is the case of a place name having another non-geographic meaning such as Nice (France), Crema, Brindisi (Italy). Geo/geo ambiguity is due to distinct locations on Earth having the same name, which occur very often for old cities of Europe having a homonymous city in the America or Australia like Rome, Paris, London, etc.

Once the geo-indexing sub-module has identified the fuzzy footprints by applying the rules in the knowledge base, it stores them into the geographic index that consists of a dual data structure:

- a direct index in which each entry is a document unique identifier and its posting list contains the fuzzy footprint of the document built during the geo-indexing phase; the direct spatial index is a B-tree organized in lexicographic order of the document identifiers, so as to optimize the retrieval of the fuzzy footprint of a document by providing as search keyword the document identifier;

- a 2D R-tree (implemented in PostGIS 1.3) in which the geographic coordinates in the footprints of all documents are organized according to their closeness on Earth so as to optimize the evaluation of punctual queries, i.e., when one specifies a location encoded by a geographic name and wants to retrieve all documents whose footprints are close to that location [30], i.e., within a maximum acceptable distance from it. The leaves of the R-tree contain the geographic coordinates gc belonging to all the footprints of the documents in the collection; each gc points to its
posting list containing its occurrences in the documents, i.e., the pairs document identifier–membership degree to the fuzzy footprint to which they belong.

On the other side, the GeoSearch module interprets queries composed of two conditions. As it can be seen in Fig. 1, there are two text fields for specifying the query: a field on the left of the GeoSearch box, where the generic content condition on the documents’ full text content representation can be specified as a set of keywords, and a field on the right of the GeoSearch box where the spatial condition can be specified as a geographic name, that is passed to the geo-indexing sub-module to identify the query footprint, and is then used to define the soft constraint on the user’s perceived distance of the documents’ footprints.

Both these two conditions are evaluated by the sub-modules named content matching module (based on the Lucene library) and spatial matching module (that is based on the novel geo-retrieval model introduced in the next section), respectively. Two distinct combinations of the content condition and spatial conditions can be chosen:

- The first combination consists in applying an asymmetric and between the content based condition, that is considered mandatory, and the spatial condition, that is considered optional. 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 evaluated for influencing the ranking of the documents (desired condition). This asymmetric combination can be tuned to enhance the influence of the spatial condition with respect to the content condition by setting a preference degree. To evaluate this combination one does not need to access the spatial index (the R-tree data structure), but just needs to access the direct spatial index providing as entries the identifiers of the documents satisfying the content based condition.

- The second combination is symmetric and compensative: in fact the degrees of satisfaction of the two conditions are aggregated linearly by specifying a relative importance between them, so that the user can give more influence to the spatial condition than to the content condition and vice versa. To evaluate in an efficient way this kind of query, one need to access the R-tree providing the locations in the query footprint as entries so as to retrieve the identifiers of the documents that satisfy the spatial conditions (in fact, the documents may not satisfy the content condition, but be relevant due to the satisfaction of the spatial condition).

3.1. The fuzzy footprint

The spatial index organizes the fuzzy footprints of documents generated during the geo-indexing phase. A fuzzy footprint of a document \(d\), \(\text{Foot}(d)\), is a fuzzy set of geographic coordinates \(gc = (lat, lon)\), where \(lat = \) latitude \(lon = \) longitude, expressed in degrees, with a membership degree \(\mu_{\text{Foot}(d)}(gc) \in [0, 1]\) representing the significance by which the geographic location \(gc\), named \(gw\), belongs to the geographic focus of the document \(d\):

\[
\text{Foot}(d) = \{(gc_1, \mu_{\text{Foot}(d)}(gc_1)), \ldots, (gc_n, \mu_{\text{Foot}(d)}(gc_n))\}
\]

where each \(gc = (lat, lon)\) and its membership degree \(\mu_{\text{Foot}(d)}(gc)\) are determined by the geo-indexing sub-module.

Each pair of coordinates \(gc=(lat,lon)\) is stored in three graphic formats for allowing the visualization with distinct applications: Gmaps format for Google maps (see Fig. 3), KML format for Google Earth, and GPX format for GPS data format environments. Further, the coordinates \(gc\) are used to access the thematic attributes of the place such as the type, population, etc. stored in the local geographic gazetteer. The local gazetteer is updated periodically with data downloaded from several on-line free sources such as Geonames [15].

4. The formal framework of the proposal

The formal framework that we adopt to define the model for documents geo-indexing is based on the use of the generalized conjunction disjunction (GCD) function [10] and is applied so as to perform a bipolar criteria evaluation like in [25]; for the recognition of geographic names we combine both favouring and disfavouring criteria.

In the following subsection we introduce the definition of the GCD function.

4.1. Generalized conjunction disjunction (GCD) function

Within the context of decision making, the generalized conjunction/disjunction (GCD) function has been defined in [10] as the weighted power mean.
Given a vector $C = [c_1, \ldots, c_m]$, with $c_i \in [0, 1]$, of satisfaction degrees of a set of criteria by an alternative, and a weighting vector $\lambda = [\lambda_1, \ldots, \lambda_m]$, with $\lambda_i \in [0, 1]$ and $\sum_{i=1,\ldots,m} \lambda_i = 1$, of distinct importance degrees of the criteria, the $GCD_p$ aggregation function is defined as follows:

$$GCD_p(C) = \left( \frac{1}{m} \sum_{i=1}^{m} \lambda_i * (c_i)^p \right)^{1/p} \quad \text{with} \quad -\infty \leq p \leq +\infty \quad p \neq 0$$

The exponent $p$ is used to adjust the logic properties of the aggregation function. By varying the value of $p$ the $GCD_p$ function can model distinct basic aggregations:

- **Simultaneity** aggregator: full conjunction, i.e., AND aggregation, is obtained with $p = -\infty$, while partial conjunction, i.e., AND–OR aggregation that linguistically could be expressed by a quantifier such as *most of*, is obtained with $-\infty < p < 1$. With these kinds of aggregators, one wants to model the situation in which the criteria *all* or *most of* are mandatory/desirable, and thus should be satisfied simultaneously.

- **Replaceability** aggregator: full disjunction, i.e., OR aggregation, is obtained with $p = +\infty$, while partial disjunction, i.e., OR–AND aggregation that linguistically could be expressed by a quantifier such as *a few*, is obtained with $1 < p < +\infty$. With these kinds of aggregators one wants to model the situation in which the criteria *a single* or *a few* are sufficient/desirable, thus the satisfaction of a *single* criterion, or of *a few* criteria, can be replaced by the satisfaction of another one, or a few other ones.

- **Neutrality** aggregator: vector product is obtained with $p = 1$. This aggregator is exactly in the middle between the AND and the OR aggregators. By choosing it, one wants a balance between simultaneity and replaceability of the criteria, i.e., a weighted average.

The semantics of a $GCD_p$ function can be captured by computing its fundamental properties of Andness for (partial) conjunctions, and Orness for (partial) disjunction aggregations. The Andness($GDC_p$) and Orness($GDC_p$) of a $GCD_p$ function are defined in the unit interval and measure the similarity between the semantics of $GDC_p$ function and the semantics of the full conjunction and disjunction. Andness($GDC_p$) and Orness($GDC_p$) are complementary indicators and satisfy the following: Andness($GDC_1$) = 1 − Orness($GDC_1$), Andness($GDC_{\infty}$) = 1, Orness($GDC_{+\infty}$) = 1, Andness($GDC_{1}$) = Orness($GDC_{1}$) = 0.5.

Therefore, a $GCD_p$ function has a mixture of conjunctive and disjunctive properties. In the case of partial conjunction, conjunctive properties predominate: Andness($GDC_{-\infty < p < 1}$) > 0.5 and Orness($GDC_{-\infty < p < 1}$) < 0.5.

Conversely, in the case of partial disjunction, disjunctive properties predominate: Andness($GDC_{1 < p < +\infty}$) < 0.5 and Orness($GDC_{1 < p < +\infty}$) > 0.5.

5. The geographic information retrieval model managing uncertainty

In this section, we describe the geographic information retrieval model proposed in this paper, that is implemented by the system Geo-Finder introduced in Section 3. The first subsection describes the geo-indexing model, while the second subsection formalizes the geo-retrieval model at the basis of the GeoSearch component.

5.1. The geo-indexing model

The geo-indexing model is defined to identify the fuzzy footprint of each document. A document is represented as a stream of tokens $\langle t \rangle$. Some terms $t$ are selected as content indexes and are stored in the content dictionary. For each entry in the content dictionary, statistical analysis is performed and information on the term frequency in the collection is stored, and, in its posting list, the documents’ identifiers in which the term appears, associated with its content significance degrees $F(d, t)$. The significance degree is defined based on statistical analysis of the document text [22]. In the posting list, the positions of the $k$ occurrences of term $t$ in the document text $\tilde{c}$, i.e., $occ_k(t, \tilde{c})$, are stored as well.

The identification of the document fuzzy footprint is achieved in three steps; at each step, distinct sets of parameterized heuristic rules $R^1$, $R^2$, and $R^3$, are evaluated, exploiting the geographic knowledge base in the local gazetteer and the lists of geographic mark up terms with words, such as city, mount, river, street, useful to disambiguate geo/non-geo terms.
Each parameterized rule has been defined manually and contains variables that are automatically instantiated with values of the geographic knowledge base: when a variable in a parameterized rule can assume \( n \) distinct values in the knowledge base, \( n \) instantiated rules are generated. By the use of parameterized rules, we achieve synthesis and readability of the heuristic knowledge and thus greater control of the geo-indexing criteria. In the Geo-Finder system, a user interface allows tuning the rules, by selecting the list of geographic mark up terms and gazetteer so as to customize the geographic indexing to a given collection (Fig. 2 depicts the Geo-Finder interface that makes it possible to configure the indexing parameters and rules). For example, if we know that the collection is written in Italian and describes Italian places, we can instantiate the parameterized rules only with the geographic names in Italy so as to reduce the number of rules actually evaluated (in Fig. 2, in the “Parameters” panel on the right one can specify the “Stopwords” file with the general stop words, and the “Geo Stopwords” file with the non-geographic terms to use, while in the “Databases” panel on the left, one can choose the “Geographic Gazetteer”).

Each set of rules acts as a filter on the input terms: only those terms, whose global satisfaction degree of the set of rules is above a threshold, are selected as input to the next step.

The geo-indexing model is implemented by the geo-indexing sub-module of the Geo-Finder system. It operates in two subsequent phases: the GeoParsing phase applies the two sets of rules, \( R_1 \) and \( R_2 \), to detect the candidate geographic names (\( gw \)) resolving geo/non-geo ambiguities. Then, the GeoCoding sub-module evaluates the third set of rules \( R_3 \) on the candidate geographic names, to identify the document fuzzy footprint \( \text{Foot}(d) \), resolving geo/geo ambiguities, and stores it in the spatial index with the document unique identifier \( d \).

**Rule set \( R_1 \):** The set of rules \( R_1 \) is aimed at realizing a pre-filtering, to reduce drastically the number of the terms on which to evaluate the set of rules \( R_2 \) that is a more costly operation. These rules are simple, like for example the following:

\[
\begin{align*}
  r_1^1 &:= \text{if Language}(d) = \text{"English"} \text{ and FirstChar}(t) \text{ is Capital letter and } t \in \text{Alphabetic string then } r_1^1(t) = 1
\end{align*}
\]

The operation of the geoparsing module receives a stream of previously selected terms as input, hereafter indicated by \( t \), and filters the candidate geographic names \( gw \) by applying the set of rules \( R_2 \).

\( R_2 \) consists of named entity recognition (NER) rules that are evaluated by applying a bipolar aggregation [9]. They exploit the local gazetteer [15] and are context dependent, where for context of a \( gw \) we mean both the geographic context and the textual context. Because a geographic name may be composed of a sequence of terms (we have estimated up to five terms), the geoparsing module works with a window on the input stream by processing five adjacent tokens.
at a time, searching for the longest geographic name matching the sequence in the local gazetteer, and then moving the window to the next token after the recognized ones. If a token or a sequence of tokens has been found in the gazetteer, the set of rules $R^2$ is evaluated.

Rule set $R^2$: There are two subsets of rules in $R^2 = R^{2+} \cup R^{2-}$, a subset of positive rules $r_i \in R^{2+}$ whose satisfaction hints to a candidate geographic name, and a subset of negative rules $r_j \in R^{2-}$ that disfavors its selection. For each $gw$, a cumulative satisfaction degree $s(gw)$ and a cumulative dissatisfaction degree $d(gw)$ are computed, that denote to what extent $gw$ is and is not a geographic name, respectively.

Once the rules are evaluated, the aggregation of the single positive (negative) satisfaction degrees of the rules is done based on a generalized conjunction disjunction (GCD) function (see Section 4). In fact, based on the values of the parameter $p$, we can model aggregations from completely replaceable (OR aggregation), where each rule can replace any other, to completely mandatory (AND aggregation), where all rules must be satisfied simultaneously [5]:

$$\begin{align*}
\text{positive}_{-}\text{hi}(gw) &= \text{GCD}_{ps}([r_1(gw) \ldots r_n(gw)]) = \left(\frac{\sum_{i=1}^{n} \lambda_i \cdot (r_i(gw))^p_s}{\sum_{i=1}^{n} \lambda_i}\right)^{1/p_s} \\
\text{negative}_{-}\text{hi}(gw) &= \text{GCD}_{pd}([r_1(gw) \ldots r_m(gw)]) = \left(\frac{\sum_{j=1}^{m} \lambda_j \cdot (r_j(gw))^p_d}{\sum_{j=1}^{m} \lambda_j}\right)^{1/p_d}
\end{align*}$$

The rules satisfaction degrees for a candidate geographic name $gw$, $r_i(gw)$ and $r_j(gw)$, assume values in $[0,1]$. $ps$ and $pd$ are set so as to define partial replaceability among rules’ satisfaction degrees. In our experiment we used $ps=pd=20$ (this value can be changed in the indexing configuration interface depicted in Fig. 2) [5].

These weights 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 the indexing phase. This way, the geographic indexing can be customized to the characteristics of a collection.

The satisfaction of a rule $r_i$ in the subset of positive rules $R^{2+}$ 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 geographic name, like, e.g., the following two rules:

$$\begin{align*}
\text{r}_1 &:= \text{"if administrative\_distr}(gw) \in \text{d} \Rightarrow r_1(gw) = 0.3" \\
\text{r}_2 &:= \text{"if } |\text{occ}_1(gw, \text{d}) - \text{occ}_1(\text{administrative\_distr}(gw), \text{d})| < \text{d} \Rightarrow r_2(gw) = 0.2"
\end{align*}$$

For instance, if $gw = \text{"San Francisco"}$ and $\text{"California"}$, the name of its administrative district, occurs in the same document $d$ at a maximum distance of $d = 3$ words, then $r_1(gw) = 0.3$, $r_2(gw) = 0.2$.

Other rules are the following:

$$\begin{align*}
\text{r}_3 &:= \text{"if } gw\_1 \in \text{prefix}(\text{type}(gw)) \Rightarrow r_3(gw) = 0.7" \\
\text{r}_4 &:= \text{"if } gw\_1 \in \text{suffix}(\text{type}(gw)) \Rightarrow r_4(gw) = 0.7"
\end{align*}$$

For instance, if $gw = \text{"Blanc"}$ is preceded by $gw\_1 = \text{"Mount"}$ that belongs to the list of prefixes {Mount, lake, river, sea, etc.} for the type of natural geographic entities to which $gw$ belongs to, then $r_3(gw) = 0.7$.

Conversely, the satisfaction of a rule $r_j$ 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:

$$\begin{align*}
\text{z}_1 &:= \text{"if } gw \in \text{Stopwords} \Rightarrow z_1(gw) = 0.8"
\end{align*}$$

For instance, $gw = \text{"Nice"}$ belongs to the stop-words list for the type cities, so $z_1(gw) = 0.8$.

The values of satisfaction of the rules can be set in the user indexing interface (see in Fig. 2, the parameters in the Geo Score window).

For each $gw$, a GeoScore$(gw, d) \in [0, 1]$ is computed based on the values positive$_{-}\text{hi}(gw)$ and negative$_{-}\text{hi}(gw)$ as a bipolar aggregation [25]:

$$\text{GeoScore}(gw, d) = \begin{cases} 
\text{positive}_{-}\text{hi}(gw) - \text{negative}_{-}\text{hi}(gw) & \text{if positive}_{-}\text{hi}(gw) > \text{negative}_{-}\text{hi}(gw) \\
0 & \text{otherwise}
\end{cases}$$

The $gw$ with GeoScore$(gw, d) \geq \tau \in [0, 1]$ are finally selected as reliable geographic names. This threshold allows restricting the fuzzy footprint of a document to reduce the possibility of identifying false positives, which is one of the major problems of geoparsing. Generally, it is better to lose some true geographic names than to select false ones.
\( \tau \) must be set based on experimental data and is also specified during the configuration of the indexing process (in Fig. 2 the value of \( \tau \) is set to 0.45 as indicated by the GeoScore threshold). The greater it is, the smaller is the possibility of selecting false geographic names.

**Rule set R**: The set of rules \( R \), used by the GeoCoding sub-module, performs the grounding or localization phase [1] that associates with the reliable geographic names their geographic locations so as to build the document fuzzy footprint (i.e., the fuzzy set of pairs of geographic coordinates \( Foot(d) \)).

For each of the selected geographic names, \( gw \), a pair of geographic coordinates \( gc = (lat, lon) \) (more generally, in case of geo/geo ambiguities, a set of geographic coordinates \( GC_{gw} = \{gc_1, ..., gc_n\} \)) is retrieved from the local gazetteer, where \( gc \) is a geocode of \( gw \), uniquely identifying a geographic place on Earth. Some properties of the place \( gc \) are also stored in the local gazetteer such as its population or administrative district.

Due to the fact that homonymous geographic names have distinct \( gc \) pairs, this phase resolves geo/geo ambiguities. To take into account the uncertainty of the disambiguation, for each geocode \( gc \), associated with a selected \( gw \), a membership degree \( \mu_{Foot(d)}(gc) \in [0, 1] \) is computed, that expresses the significance 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 \( x'_i(gc) \in [0, 1] \) computed by taking into account the context of the geographic names.

For example, the following rule \( x'_1 \) considers the textual context in the document of the occurrences of a geographic name \( gw \) in order to increase the membership degrees of all its geocodes in the \( Foot(d) \): the frequency of a \( gw \) in a document \( d \), \( F(d, gw) \), increases the membership degrees in the footprint proportional to the degree by which \( gw \) was recognized as a candidate geographic name:

\[
x'_1 := "If \ gc \in GC_{gw} \Rightarrow x'_1(gc) = \mu_{significant}(F(d, gw) \ast GeoScore(gc, d))"
\]

where \( \mu_{significant} \) is a monotonic non-decreasing membership function.

The following rules are aimed at resolving geo/geo ambiguities, that is, they discriminate the membership degrees in \( Foot(d) \) of the geocodes \( gc \in GC_{gw} \) of a \( gw \).

For instance, Rule \( x'_2 \) evaluates the geographic context of the geographic names \( gw \) in a document to disambiguate their location: given a geocode \( gc \in GC_{gw} \) the membership degree of \( cg \) in \( Foot(d) \) is increased if there exists another \( gw_k \) in \( d \) that is the name of the administrative district of \( gw \) located in \( gc \) or it belongs to the same administrative district of \( gw \):

\[
x'_2 := "If \ gc \in GC_{gw} \land gw_k \in d \land gw_k \neq gw \land gw_k = administrative_distr(gc) \lor \exists gc_k \in GC_{gw_k} |
\administrative_distr(gc_k) = administrative_distr(gc)
\Rightarrow x'_2(gc) = \left( \sum_{gw_k \in d} F(d, gw_k) / MaxF \right) ^*"
\]

where \( MaxF \) is a normalization parameter.

The following rule \( x'_3 \) discriminates the geocode membership degrees in \( Foot(d) \) based on the geographic distance \( dist \) between the geographic names in the document text, so that co-occurring geographic names, with geocodes geographically close to each other get a greater membership value:

\[
x'_3 := "If \ gc \in GC_{gw} \land gw_k \in d \land gw_k \neq gw \Rightarrow x'_3(gc) = \left( \sum_{gc_k \in GC_{gw_k}} \mu_{near}(dist(gc, gc_k)) \right) / N ^*"
\]

with \( \mu_{near} \) being a non-increasing function defined in the implementation as follows:

\[
\mu_{near}(x) = \begin{cases} 
1 & \text{if } x < \text{AvgDist}(d) \\
(x)^{1/3} & \text{otherwise}
\end{cases}
\]
where $\text{AvgDist}(d) = (\sum_{gc_i \in GC_{gw}} \frac{\text{dist}(gc_i, gc_j))}{N}$ is the average geographic distance $\text{dist}$ for document $d$ (it is a value relative to the document footprint). $\text{dist}$ is a great circle approximation of the actual distance between the two spherical coordinates $i$ and $j$.

Rule $x_4'$ disambiguates the geocodes $gc \in GC_{gw}$ of a geographic name $gw$ by taking into account the population attribute $\text{population}(gc)$: the value of the population is used to favour the membership degree of big cities (with a population above the average population) with respect to small ones in the footprint:

$$x_4' := If \ gc \in GC_{gw} \Rightarrow x_4'(gc) = \mu_{\text{big}}(\text{population}(gc)/\text{PopAvg})^*$$

where $\mu_{\text{big}} : N \rightarrow [0, 1]$ is a non-decreasing function defined in the implementation as follows:

$$\mu_{\text{big}}(x) = \begin{cases} (x)^{1/3} & \text{if } x < \text{PopAvg} \\ 1 & \text{otherwise} \end{cases}$$

where $\text{PopAvg}$ is the average of the population values in the gazetteer.

Finally, all rules in the set $R^2$ are aggregated based on a $\text{GCD}_p$ aggregation function, 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 as the frequency of activation of each rule on a training set, and the parameter $p=1$, i.e., neutral aggregation:

$$\text{geo_significance}(gc) = \text{GCD}_p([x_1'(gc) \ldots x_m'(gc)]) = \left(\sum_{i=1}^{m} \lambda_i ^* (x_i'(gc))^p \right)^{1/p} \quad (4)$$

Therefore, at the end of this step a geographic name $gw$ may have several geocodes with distinct geographic significance degree $\text{geo_significance}(gc)$: we select the $gc$ of a geographic name $gw$ with greatest membership value:

$$gc \in \text{Foot}(d) \ with \ \mu_{\text{Foot}(d)}(gc) = \text{geo_significance}(cg)|gc = \text{ArgMax}_{i \in GW}(\text{geo_significance}(i)) \quad (5)$$

Finally, a minimum threshold $\Phi \geq 0$ on $\mu_{\text{Foot}(d)}(i)$ may be specified to restrict the extent of the footprint of a document (in Fig. 2 $\Phi$ is set equal to 0.5 as indicated by the value of $\text{GeoRefValue\ threshold}$). One can also specify a maximum number of name places in a footprint of a document (in Fig. 2, this maximum number is set to 15 as indicated by $\text{GeoWord\ number}$).

$$\text{Foot}(d) = \{cg_1, \mu_{\text{Foot}(d)}(gc_1)), \ldots, (cg_n, \mu_{\text{Foot}(d)}(gc_n))|\mu_{\text{Foot}(d)}(gc_j) \geq \Phi$$

$$\wedge n < \text{GeoWord\ number} \wedge gc_j \in GC_{gw} | \text{GeoScore}(gw_k, d) \geq \tau \}$$

with $\Phi$ and $\tau$ in $[0,1]$ and $\text{GeoWord\ number} \in N$.

Fig. 3 shows the Footprint extracted from a sample paper. The map was generated by the $\text{Mapping\ module}$ of GeoFinder that uses distinct graphic formats for representing the document's footprint, such as Gmaps for Google maps (in Fig. 3), KML for Google Earth, and GPX for GPS data format environments.

5.2. The geo-retrieval model

The geo-retrieval model parses a user query that consists of two selection 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 content “close” to the specified place names. These two conditions are evaluated by their partial matching functions computing two scores in $[0,1]$, the content retrieval status value, indicated by $RSV_{\text{content}}(d)$, and the geographic retrieval value, indicated by $GRV_{\text{closeness}}(d)$, respectively. These partial scores are finally combined to compute the global retrieval status value with respect to the whole query $q$, indicated by $RSV_q(d)$, by applying an aggregation function that can be chosen by the user between two possible ones:

- “and possibly” asymmetric aggregation between a mandatory condition, the content one, and a desired condition, the spatial one. The user can express a preference of the second condition with respect to the first one.
- “average” compensative aggregation of the two weighted conditions having distinct relative importance.
In the following we formalize the definitions of the two partial matching functions and of the two possible combinations of the partial scores.

5.2.1. Evaluation of the content based condition

To compute the satisfaction of the content condition in the query, the retrieval functions of the Lucene library are used [22], based on the vector space retrieval model, in which the content keywords are interpreted as identifying an ideal document in the document space. For each keyword in the query the full-text inverted index is accessed and then the inner product is used to measure the similarity between the document’s vector and the content keywords’ query vector; the inner product allows not penalizing long documents with respect to short ones. $RSV_{content}(\overline{d})$ is thus computed.

5.2.2. Evaluation of the spatial condition

To evaluate the spatial condition, firstly the fuzzy footprint $Foot(q)$ of the geographic names in the query $q$ is identified by applying the GeoParsing and GeoCoding rules, described in the previous subsection.

Secondly, the fuzzy footprints of the documents $d$, $Foot(d)$ that are likely to satisfy the query are retrieved by accessing the spatial index: in the case of “and possibly” combination the identifiers of the documents satisfying the content
condition are used as entries in the direct spatial index, i.e., the B-tree, while in the case of “average” combination the geographic coordinates in $Foot(q)$ are used as entries in the R-tree spatial index.

In the current implementation of the system Geo-Finder, 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 $\mu_{close}$ which models the concept “close” as a user’s context dependent soft constraint.

Future extensions of this step will allow the specification of other spatial conditions that can also be represented as context depend soft constraints as proposed in [6].

The matching function $\mu_{close}$ 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:

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

(6)

where $\mu_{Foot}(q)(i)$ and $\mu_{Foot}(q)(j)$ are the membership degrees of the $i$-th and $j$-th pairs of geographic coordinates latitude and longitude in the fuzzy sets $Foot(d)$ and $Foot(q)$, i.e., the significance degrees in representing the geographic focus of the document and the query, respectively.

dist$(i, j)$ is a great circle approximation of the actual distance between the two spherical coordinates $i$ and $j$.

$qscope(x)$ function modifies the geographic distance so as to model the user perceived distance as follows:

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

(7)

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.

$\delta$ and $k$ have default values, and do not need the user’s explicit input. Nevertheless, the user can specify their desired values if he wants it.

The parameter $\delta$ 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 $\delta$ can adapt the evaluation of the spatial condition “close” to the user perception, thus, modelling strict or relaxed interpretations of the “closeness” surroundings of a point. The higher the $\delta$, the greater is the surrounding.

In a dynamic context where the user is moving $\delta$ could be set by taking into account the user’s speed as proposed in [4].

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$(Foot(q))$, 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 $\delta$ see in Fig. 2 the query scope settings):

- a small scope, defined with $k = 5, \delta = 3$ km, 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;
- a meso scope, defined with $k = 4, \delta = 50$ km, when MaxDist$(Foot(q))$ covers the area of either a region or a small nation like Belgium (in this case $Foot(q)$ could vary approximately between 0 and a few hundred kilometres): with this setting one can retrieve documents within a distance from the query of 50 km to a few thousands kilometres;
- a large scope, defined with $k = 3, \delta = 1000$ km, when MaxDist$(Foot(q))$ covers the area of a medium nation such as France (in this case $Foot(q)$ could vary approximately between 0 and a few thousand kilometres): with this setting one can retrieve documents within a distance from the query of one thousand to a few ten thousands kilometres;
- a full scope, defined with $k = 3, \delta = 10000$ km when MaxDist$(Foot(q))$ covers the area of a big nation such as Russia or of a continent (in this case $Foot(q)$ could vary approximately between 0 and more than 10000 km). This allows considering the whole globe.
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 \(d = 50\) 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.

On the other side, a query with footprint in Bergamo, Dalmine (10 km from Bergamo in Milano direction) and the same meso scope will retrieve documents at a maximum distance from the query footprint of 90 km: given the same example above, the query will retrieve just the document in “Milano” and not the one in “Lugano”.

5.2.3. “And possibly” combination

The evaluation of the asymmetric combination of the content condition and the spatial condition is implemented by the sequential evaluation of the two conditions: first the content condition is evaluated, and, only for the documents that satisfy it, i.e., in the case in which \(RSV_{\text{content}}(d) > 0\), also the spatial condition is evaluated by computing the geographic retrieval value, \(GRV_{\text{foot}}(d) \in [0, 1]\). This sequential evaluation allows reducing the number of documents on which evaluating the most costly spatial condition. The footprints of the documents \(d\) satisfying the content condition (\(RSV_{\text{content}}(d) > 0\)) are retrieved from the direct spatial index by giving as entries the documents’ unique identifiers.

Finally, the global retrieval status value of the query, \(RSV_q(d)\), is obtained by the asymmetric combination of the two partial scores \([9,38]\) by applying the following aggregation:

\[
RSV_q(d) = RSV_{\text{content}}(d) \text{ and possibly} \ GRV_{\text{closeness}}(d) = RSV_{\text{content}}(d) \times \max((1 - \alpha), GRV_{\text{closeness}}(d))
\]

\(\alpha\) specifies the user’s preference of the spatial condition with respect to 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 \(RSV_{\text{content}}(d)\). When \(\alpha = 1\) the two conditions are both mandatory: this means that the geographic retrieval value \(GRV_{\text{closeness}}(d)\) has the same relevance of the content retrieval status value \(RSV_{\text{content}}(d)\). In this case the aggregation reduces to the product, i.e., the “fuzzy Anding” of the two relevance scores. Intermediate values of \(\alpha\) in (0,1) demands for an asymmetric combination. The value \((1 - \alpha)\) guarantees a minimum satisfaction level for \(GRV_{\text{closeness}}(d)\) 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.

5.2.4. Average combination

When specifying this combination, the content condition and the spatial condition are evaluated by parallel independent processes, by applying their partial matching functions. In this case each pairs of geographic coordinates in the footprint of the query is used as entry in the spatial R-tree.

Finally, the global retrieval status value of the query, \(RSV_q(d)\), is obtained by the linear combination of the two partial scores \([9,38]\), in which one can set an importance degree \(\alpha \in [0, 1]\) of a condition with respect to the other defined as follows:

\[
RSV_q(d) = RSV_{\text{content}}(d) \text{ average} \ GRV_{\text{closeness}}(d) = (1 - \alpha) \times RSV_{\text{content}}(d) + \alpha \times GRV_{\text{closeness}}(d)
\]

The main difference of this definition of the global RSV with respect to the definition given by the asymmetric combination defined in (8) is that the two conditions compensate one another, while with the previous definition, the satisfaction of the content condition was necessary to retrieve a document. In (9) 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 \(\alpha\) allow varying the trade-off between the influences of the two conditions.

6. Evaluation of Geo-Finder system

Fig. 4 depicts the GeoSearch user interface of the Geo-Finder system. At the top, there are two text forms for submitting the content (left) and spatial (right) query condition, plus a menu for selecting the desired combination of the two conditions. One can also perform an advanced search by both indicating the textual fields of documents that must match the content condition, and the desired query scope of the spatial condition (full, large, meso and small, from
Fig. 4. GeoSearch graphic user interface of Geo-Finder: at the top the two query fields for specifying the content condition="See side" and spatial condition="Sao Paulo") combined by “and possibly”; below the two results panels. On the right side the bi-dimensional relevance graph, on the left side the ranked list.

left to right). Notice that one could change the definitions of the query scopes by going to the indexing configuration interface in Fig. 2.

At the Bottom, on the right panel, the bi-dimensional relevance graph is depicted, in which each point corresponds to a retrieved document. The origin identifies the query. The document’s X coordinate and Y coordinate are its relevance degrees with respect to the spatial condition and the textual content condition respectively.

On the left panel, the ordered list of retrieved documents is shown. In Fig. 4, the ranking is determined by the “and possibly” combination, defined in Eq. (8), with equal preference for the mandatory and spatial conditions, i.e., AND. 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.

The system has undergone a user evaluation based on the GeoClef 2006 topics reported in Table 1.

For expressing the needs of the GeoClef 2006 topics, for each of them we defined a query consisting of a set of terms, for specifying the content condition (third column), and a set of geographic names, for specifying the spatial condition (fourth column). The information needs of the GC040 topic (line 15 of Table 1) does not require specifying a “proper” spatial condition, because any geographic place of type “city” that is close to any place of type “volcanoes” is relevant to the topic. Due to the fact that Geo-Finder is unable to select documents about geographic places of a desired typology, this topic was associated with a query just consisting of the content condition.

In order to generate a collection of documents which are likely to be relevant to the topics we submitted to Google 1 25 queries defined by the lists of the terms that appear in the third and fourth columns of Table 1, in the order reported by concatenating the two columns.

A collection of 250 web pages was generated with the top 10 ranked pages retrieved by Google as a result of each of the 25 queries. Notice that Google interprets a query consisting of a list of terms as an AND/OR query; if it does not retrieve a sufficient number of web pages that contain all the query terms it also retrieves the web pages that contain less query terms.

---

1 The web pages were retrieved by Google on January 2010.
Table 1
Topics of GeoClef 2006 test: from left to right the topic Identifier, the title, the query content condition and the query spatial condition. The queries submitted to Google consist of the list of terms appearing in both the content condition and the spatial condition, in the same order.

<table>
<thead>
<tr>
<th>Query Id.</th>
<th>Topic title</th>
<th>Google and Geo-Finder query</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GC026</td>
<td>Wine regions around rivers in Europe</td>
<td>Wine regions rivers Europe</td>
</tr>
<tr>
<td>2. GC027</td>
<td>Cities within 100km from Frankfurt</td>
<td>Cities 100km Frankfurt</td>
</tr>
<tr>
<td>3. GC028</td>
<td>Snowstorms in North America</td>
<td>Snowstorms North America</td>
</tr>
<tr>
<td>4. GC029</td>
<td>Diamond trade in Angola and South Africa</td>
<td>Diamond trade Angola South Africa</td>
</tr>
<tr>
<td>5. GC030</td>
<td>Car bombings near Madrid</td>
<td>Car bombings Madrid</td>
</tr>
<tr>
<td>6. GC031</td>
<td>Combats and embargo in the northern part of Iraq</td>
<td>Combats embargo North Iraq</td>
</tr>
<tr>
<td>7. GC032</td>
<td>Independence movement in Quebec</td>
<td>Independence movement Quebec</td>
</tr>
<tr>
<td>8. GC033</td>
<td>International sports competitions in the Ruhr area</td>
<td>International sports competitions Ruhr</td>
</tr>
<tr>
<td>9. GC034</td>
<td>Malaria in the tropics</td>
<td>Malaria Tropics</td>
</tr>
<tr>
<td>10. GC035</td>
<td>Credits to the former Eastern Bloc</td>
<td>Credits Eastern Bloc</td>
</tr>
<tr>
<td>11. GC036</td>
<td>Automotive industry around the Sea of Japan</td>
<td>Automotive industry Japan</td>
</tr>
<tr>
<td>12. GC037</td>
<td>Archaeology in the Middle East</td>
<td>Archeology Middle East</td>
</tr>
<tr>
<td>13. GC038</td>
<td>Solar or lunar eclipse in Southeast Asia</td>
<td>Solar lunar eclipse Southeast Asia</td>
</tr>
<tr>
<td>14. GC039</td>
<td>Russian troops in the southern Caucasus</td>
<td>Russian troops Southern Caucasus</td>
</tr>
<tr>
<td>15. GC040</td>
<td>Cities near active volcanoes</td>
<td>Cities active volcanoes</td>
</tr>
<tr>
<td>16. GC041</td>
<td>Shipwrecks in the Atlantic Ocean</td>
<td>Shipwrecks Atlantic Ocean</td>
</tr>
<tr>
<td>17. GC042</td>
<td>Regional elections in Northern Germany</td>
<td>Regional elections Northern Germany</td>
</tr>
<tr>
<td>18. GC043</td>
<td>Scientific research in New England Universities</td>
<td>Scientific research Universities New England</td>
</tr>
<tr>
<td>19. GC044</td>
<td>Arms sales in former Yugoslavia</td>
<td>Arms sale Yugoslavia</td>
</tr>
<tr>
<td>20. GC045</td>
<td>Tourism in Northeast Brazil</td>
<td>Tourism Northeast Brazil</td>
</tr>
<tr>
<td>21. GC046</td>
<td>Forest fires in Northern Portugal</td>
<td>Forest fires Northern Portugal</td>
</tr>
<tr>
<td>22. GC047</td>
<td>Champions League games near the Mediterranean</td>
<td>Champions League games Mediterranean</td>
</tr>
<tr>
<td>23. GC048</td>
<td>Fishing in Newfoundland and Greenland</td>
<td>Fishing Newfoundland Greenland</td>
</tr>
<tr>
<td>24. GC049</td>
<td>ETA in France</td>
<td>ETA France</td>
</tr>
<tr>
<td>25. GC050</td>
<td>Cities along the Danube and the Rhine</td>
<td>Cities Danube River Rhine River</td>
</tr>
</tbody>
</table>

So it is not straightforward that the top 10 ranked pages by Google are all relevant to a user because it might happen that some of them do not contain the geographic names specified in the query. For this reason we asked two users to assess the relevance of these web pages with respect to the topics. The evaluation of the relevance of each of the 250 web pages to the topics was manually assessed by each user who specified if the web page was relevant or not to the query and the precision of Google at the 10 recall levels was computed for each query and each user and then averaged for the two users.

In order to evaluate Geo-Finder, first of all, the 250 documents were indexed by the Geo-Finder indexing module to generate both the content indexes and spatial indexes.

Then, the queries in Table 1 were submitted and evaluated by Geo-Finder.

Firstly, the queries were submitted by combining the two conditions with “And possibly” with maximum preference ($\alpha = 1$), which means a pure ANDing of the two conditions; successively, by combining the conditions with “Average” and equal preferences ($\alpha = 0.5$). In both the cases the default query scope full was used. These settings of the preferences and of the query scope were chosen so as to submit to Geo-Finder queries with a semantics as far as possible similar to the queries submitted to Google.

We observed an overall better performance of Geo-Finder with respect to Google: the average precision at 10 top retrieved documents is 0.7 for Google and 0.77 for Geo-Finder with both combinations of the conditions in the query. A better performance of Geo-Finder when evaluating the ANDing of the conditions than the Average combination was observed.

Figs. 5 and 6 report the difference between the values of the precision at top 5 retrieved documents (denoted by 5-P), at top 10 retrieved documents (denoted by 10-P), and the average precision for all the 10 levels of recall (denoted by $\{1,10\}$-P) by Geo-Finder with respect to Google for the 25 queries with the “And” and the “Average” combination of the two conditions, respectively.
The first observation we can make by looking at the average precision for all the 10 recall levels in the two figures is that, apart for two queries (query #5 is answered better by Google when using “And” and by Geo-Finder when using “Average” combination, and query #20 is answered better by Geo-Finder when using “And” and by Google when using “Average”), the positive/negative performance of Geo-Finder with respect to Google is the same independently from the used combination operator. Nevertheless a difference depending on the combination operator can be appreciated by a quantitative comparison of the precision at top 5 and at top 10 retrieved documents.

In Fig. 5 it can be observed that Geo-FinderAnd is more precise than Google in ranking the first five documents in all the 25 queries, it is greater or at least as precise as Google in 72% of the queries in ranking the first 10 documents, and finally its average precision over all the 10 recall levels is at least as that of Google in 68% of the queries.

In Fig. 6 it can be observed that Geo-FinderAverage is greater or at least as precise as Google in 64% of the queries in ranking the first five documents, in 68% of the queries in ranking the first 10 documents, and its average precision over all the 10 recall levels is at least as that of Google in 68% of the queries.

From these results we observe that the “And” combination allows achieving a greater precision at top 5 and at top 10 ranked positions of the retrieved list with respect to the “Average” combination.

Further, we wanted to analyse how the precision of Geo-Finder varies depending on the user context, i.e., by distinct query scopes. Fig. 7 illustrates the precision values at the 10 standard recall levels for GeoFinderAnd when considering the four types of query scope, full, large, meso and small and for Google.

Fig. 8 has the same meaning of Fig. 7 but in this case the query evaluated by Geo-Finder used the “Average” combination.

In both Figs. 7 and 8 it can be seen that the precision increases on almost all the recall levels when reducing the query scope. Further, Geo-Finder outperforms Google over all recall levels when the query scope is meso and small, which means that the possibility to restrict the geographic domain of interest allows improving the ranking of the relevant documents.
Fig. 6. Graph representing the increment/decrement of the values of precision at top 5 ranked documents, at top 10 ranked documents, and averagely over all recall levels of Geo-Finder\textsubscript{Average} with respect to Google.

Fig. 7. Precision at the 10 Recall levels of Geo-Finder\textsubscript{And} with distinct query scopes, and Google.

With both combination operators, the meso query scope produces the best results over all other query scopes, even with respect to the small query scope, that never produces greater precision values than the meso scope for any level of recall. This means that when restricting too much the query scope it becomes difficult to capture the geographic area of interest to the user.
7. Conclusions

The contributions of this proposal with respect to current practice in geographic information retrieval are several. First of all, a fuzzy footprint is computed to represent the geo-reference focus of a textual document thus by taking into account the uncertainty of the indexing: a place on Earth can be associated with a document to a degree, expressing the significance in representing the document geographic focus. To this aim multiple bipolar criteria are evaluated during the geo-indexing process.

Second, flexible queries can be submitted to the retrieval module of Geo-Finder system expressing two possible semantics: either a mandatory content condition, and a desired soft context dependent spatial condition, or a compensative aggregation of the two conditions with a relative preference between them. The spatial condition is defined so as to take into account a user’s perceived distance measure.

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 modelling the user’s context which determines the perceived distance when evaluating the spatial query condition.

References
