Alma Mater Studiorum — Università di Bologna

DOTTORATO DI RICERCA IN COMPUTER SCIENCE AND ENGINEERING

Ciclo: XXIX

Settore Concorsuale di Afferenza: 01/B1
Settore Scientifico Disciplinare: INF/01

# GEIR: a full-fledged **G**eographically **E**nhanced **I**nformation **R**etrieval solution

Presentata da: Yisleidy Linares Zaila

Coordinatore Dottorato:
Paolo Ciaccia

Relatore:
Danilo Montesi

Esame finale anno 2017

# Abstract

With the development of search engines (e.g. Google, Bing, Yahoo, etc.), people is ambitiously expecting higher quality and improvements of current technologies. Bringing human intelligence features to these tools, like the ability to find implicit information through semantics, is one of the must prominent research lines in Computer Science. Information semantics is a very wide concept, as wide as the human capability to interpret, in particular, the analysis of geographical semantics gives the possibility to associate information with a place. It is estimated that more than 70% of all information in the world has some kind of geographic features [48]. In 2012, Ed Parsons, a GeoSpatial Technologist from Google, reported that between 30% and 40% of the user queries at Google search engine contain geographic references [16].

This thesis addresses the field of geographic information extraction and retrieval in unstructured texts. This process includes the identification of spatial features in textual documents, the data indexing, the manipulation of the relevance of the identified geographic entities and the multi-criteria retrieval according to the thematic and geographic information.

The main contributions of this work include a custom geographic knowledge base, built from the combination of GeoNames and WordNet; a Natural Language Processing and knowledge based heuristics for Toponym Recognition and Toponym Disambiguation; and a geographic relevance weighting model that supports non-spatial indexing and simple ranking combination approaches. The validity of each one of these components is supported by practical experiments that show their effectiveness in different scenarios and their alignment with state of the art solutions.

In addition, it also constitutes a main contribution of this work GEIR, a general purpose GIR framework that includes the implementations of the above described components and brings the possibility of implementing new ones and test their performance within an end to end GIR system.

# Acknowledgements

First and foremost I want to thank my advisor Prof. Danilo Montesi for giving me the amazing opportunity of coming from very far away to become one of his Ph.D. students, also for his guidance and patience along these three years.

I would also like to thank my reviewers, specially Prof. Ross Purves and Prof. Christopher Jones, for the great care they put in reading my thesis and for the excellent advices they gave me.

I want to express my gratefulness as well to my fellow Ph.D. colleagues for their great company in personal and professional times.

I want to thank my family for their constant encouraging thoughts, and specially my parents, to whom I would never be able to express enough gratitude for all the efforts and sacrifices made on my behalf.

A very special thanks goes to Abel, for his total love, having you by my side at each step of the way has been a real blessing.

Finally, thank you God for letting me through all the challenging and difficult moments!

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

As part of the Information Retrieval (IR) discipline, the Geographic IR (GIR) inherits most concepts, theories and techniques from its ancestor. Thus, the basic knowledge of conventional IR is rather beneficial as a background for the understanding of a GIR system. According to Larson [57], a generic information retrieval system is composed by five elements:

1. **Indexing**: This element deals with the organization of the documents of a retrieval system for obtaining an efficient access. In IR systems, indexing is derived from the contents, such as keyword extraction. IR can also index terms from a controlled vocabulary, which is a pre-selected list of words that can be used to describe any document in the collection. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query.

2. **Retrieval Models**: A model of information retrieval predicts and explains what a user will find relevant given the user query. Some of IR models are: the boolean model, where documents are retrieved but are not ranked, simply classified as relevant and not relevant; the vector space model, where query and documents are represented as vectors embedded in a high dimensional Euclidean space; the probabilistic model where term weighting are based on probability theory; and Google's page rank model which used a radically different approach for ranking the search results based on the hyperlink structure of the web.

3. **Matching and Retrieving**: This element refers to the connection between the information needed by the user and the documents in the collection. The system finds information to given criteria by matching documents against user queries. A retrieved document might not be the ideal, but it will be close enough from the system's point of view.

4. **Relevance and Ordering**: Fourth and fifth elements concern to the display of information. Each retrieval system has a criteria for the items retrieved. Relevance is the reason why the items are retrieved. An item that deals with that exact topic will be considered extremely relevant. The retrieved items are ranked in a particular order, the most relevant items at the top of the list and the less relevant at the end.

5. **Query Languages**: This component deals with the query mechanism of the retrieval system. In IR systems, queries are composed using natural language. The system needs to be capable to interpret the query in a syntactical and semantical sense in order to fulfil the user needs.

GIR systems require a more complex data representation, information retrieval model and system architecture than traditional IR systems. Currently, one of the most important challenges in GIR systems is the provision of detailed semantic information that allows to analyze the spatial relationship among geographic entities present in unstructured texts.

The concept of GIR, also introduced by Larson [57], is defined as the process related to providing access to geo-referenced information sources. It is about the theories and technologies to enable traditional Information Retrieval (IR) systems to better answer users queries bearing some sorts of geographical semantics. The main task of GIR sysems is the definition of techniques to build an application system that could well index, retrieve and browse the geo-referenced information. The following are some examples of the matters that GIR system aim to solve [69]:

- **non-geographic subjects restricted to a place**: when the non-geographic themes only can take place in a specific location (e.g., *Redentore Festival in Venice*)

- **geographic subjects with non-geographic restrictions**: when the geographic themes can be found everywhere (e.g., *lakes among mountains*)

- **geographic subjects restricted to a place**: when the geographic themes can be only found in a specific location (e.g., *rivers in Italy*)

- **non-geographic subjects associated with a place**: when a non-geographic theme is found in a location explicitly specified but it also can be associated to other places (e.g., *independence of Cuba*)

- **non-geographic subjects that are a complex function of a place**: when the geographic information is implicitly included into a non-geographic theme (e.g., *European Conference on Information Retrieval*)

- **geographical relations among places**: it refers to the different relationships that can be defined among locations (e.g., *how is San Vitale related to Bologna? Is it inside?*)

From these questions, we can identify four basic features relevant to geographically oriented questions:

- **place-names**: a place-name, also known as toponym, is the name given to or held by a geographical location. (e.g., *Venice, Italy, Cuba, San Vitale, Bologna, etc.*)

- **geographical relations**: it contains all possible relationships that can be defined among locations (e.g., *south of, close to, inside of, etc.*)

- **geographical concepts**: it includes all the different types of geographic elements (e.g., *lakes, mountains, rivers, etc.*)

- **geographical adjectives or demonyms**: this group contains the words that identify residents or natives of particular places. Each of these words are derived from the name of its corresponding place (e.g., *European, etc.*)

All these geographical terms have an active participation in the development of GIR models providing the basis for facing very important problems in GIR such as: toponym resolution, geographic footprint of documents and queries, geographic ranking of documents according to a specific query and geographic similarity function for comparing documents and queries. There are some approaches focused on this direction [28,48,56] being GeoCLEF[1] in 2005, the first major campaign to promote the development of GIR systems. This campaign has been the inspiration for important advances [26, 60, 73], however GIR is still a relatively young branch where the main challenge is to satisfy user needs properly through the analysis of geographical features present in their queries and in the information source.

## 1.1 Problem Statement

The assumption of this thesis is that a traditional Information Retrieval (IR) search engine can be improved by adding geographical analysis. This thesis poses the question whether the geographical analysis of documents and user queries can be exploited to improve information retrieval results in terms of accuracy and relevance. Currently, there are several approaches dealing with the geographical analysis in unstructured texts. Some of them constitute full GIR solutions [24,48,73,80] while others aim to solve particular problems that need to be handled for developing a GIR solution, such as toponym resolution [1,4,5,19,23,39,52], geographic query processing [27,29,38,112] or geographic similarity and ranking [10,32,47]. Nowadays, among other challenging problems, researchers in GIR have to face two tasks:

1. If the research is about the resolution of one specific problem in GIR, there is a need to develop a whole GIR system in order to evaluate the performance of that specific proposed approach. For example, how to assess the effect of a new toponym resolution technique over the geographic ranking process if there is not any GIR solution available allowing to insert the new technique in its architecture?

2. Comparing any new approach with other existing ones is currently, almost impossible. Very few techniques are available for downloading or are described with enough detail in order to reproduce them.

This thesis aims to solve these issues by the development of a GIR framework, called GEIR (Geographic Enhanced Information Retrieval). In IR, there are different frameworks that provide the basis for easily developing new approaches and to compare them with standard IR techniques (i.e., Terrier[2], Lucene[3]). To the best of our knowledge, there is no any similar framework oriented to GIR. Therefore, this thesis presents GEIR as an extension of Terrier IR framework, which not only provides the tools for easily developing approaches related to

[1] http://clef-campaign.org

[2] http://terrier.org/

[3] http://lucene.apache.org/core/

the geographic analysis in unstructured texts, but also includes a whole GIR solution as a starting point for the development of new tools.

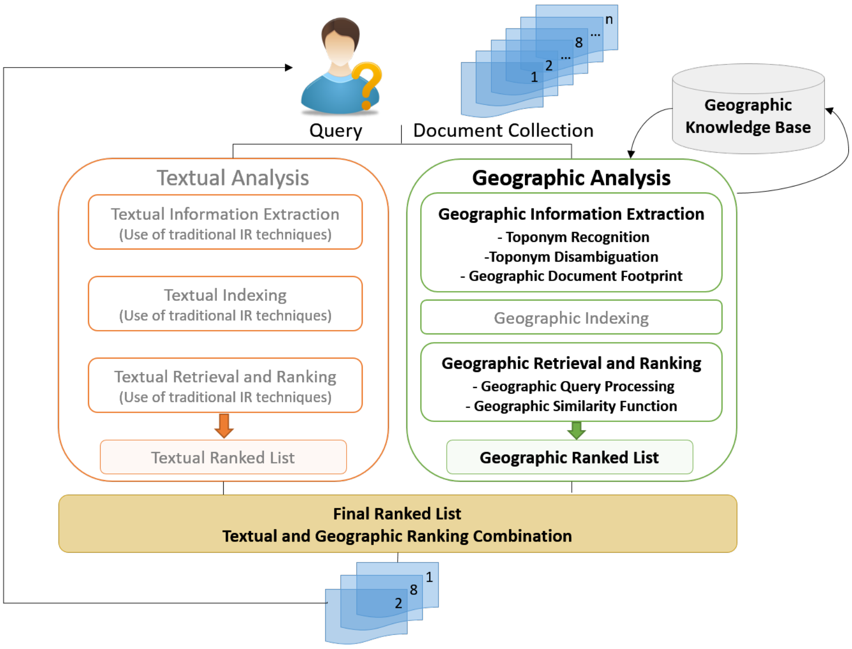
Query
Document Collection
1
2
8
n
Geographic Knowledge Base
Textual Analysis
Textual Information Extraction (Use of traditional IR techniques)
Textual Indexing (Use of traditional IR techniques)
Textual Retrieval and Ranking (Use of traditional IR techniques)
Textual Ranked List
Geographic Analysis
Geographic Information Extraction
- Toponym Recognition
-Toponym Disambiguation
- Geographic Document Footprint
Geographic Indexing
Geographic Retrieval and Ranking
- Geographic Query Processing
- Geographic Similarity Function
Geographic Ranked List
Final Ranked List
Textual and Geographic Ranking Combination

Figure 1.1: General GIR system architecture. In bold, the tasks tackled in this thesis.

A general GIR system can be structured as shown in Figure 1.1. This is the structure considered in this work. Given a query and a collection of documents the information is processed from both, textual and geographic points of view. The textual analysis employs traditional IR techniques for generating a textual ranked list of documents that responds to the textual information needs, while the geographic analysis includes the following tasks which constitute the research lines handled in this thesis:

- **toponym recognition**, which refers to the identification of geographic features in documents

- **toponym disambiguation**, that is intended to assign the corresponding geographic coordinates to each previously recognized toponym

- **geographic document footprint**, which manages the relevance of the geographic features present in documents as well as its representation

- **geographic query processing**, that includes the analysis of the geographic information present in the user query which is slightly different from the geographic analysis of documents because the query length is quite smaller than document length

- **geographic similarity function**, which defines how to compare documents and queries according to their geographic information providing a geographic ranked list of documents that responds to the geographic information needs

- **textual and geographic ranking combination** for producing a final ranking list of documents that attempts to better satisfy the user needs.

As an important remark we notice that the a very common element of GIR systems has been left out of the proposed solution.

## 1.2 Objectives

This section formally defines the objectives this thesis aims to achieve in order to solve the above stated problem.

### General Objectives

1. The analysis of the geographical information in unstructured texts for handling GIR tasks.
2. The development of a framework oriented to GIR which provides the tools for easily developing and assessing new approaches dedicated to the geographic analysis in texts.

**Specific Objectives** The first general objective is divided in the following specific ones:

- toponym resolution: Recognition and disambiguation of geographic references.
- relevance ranking: Use of the analyzed geographic information for obtaining more accurate results in the retrieval process.

The toponym resolution objective is further split into two subordinated objectives: i) Toponym Recognition (TR) and ii) Toponym Disambiguation (TD). The former is the process of identifying geographic references in unstructured text. It is intended to determine if a word or sequence of words is a geographic reference or not. The latter is the process of assigning to each geographic reference unique geographic coordinates.

The relevance ranking objective is, in turn, split into three subordinated objectives: i) definition of the geographic footprint of documents/queries, ii) definition of a geographic similarity function and iii) combination of geographic and textual ranking. The geographic footprint of documents/queries allows to represent the geographic information present in documents/queries and to quantify its importance. The geographic similarity function describes how to compare the geographic information detected in documents and queries in order to generate a ranked list of documents according to their geography. The combination of geographic and textual ranking defines how to combine the results obtained from the textual and geographical analyses in order to generate a more accurate final ranked list of documents.

On the other hand, the second general objective introduces the following specific objectives:

- Design of a modular architecture that allows to implement new approaches and to easily test their performance within a whole GIR solution.

- Implementation of a whole GIR solution that constitutes a baseline for further developments.

The implementation of a whole GIR solution is directly obtained from the first general objective. Each proposed GIR component (toponym resolution, document processing, query processing, geographic ranking, textual and geographic ranking combination) requires to be implemented according to a proper architecture design that satisfies the above desired features.

## 1.3 Thesis Contribution

In synthesis, the contributions of this thesis to the GIR research area are:

1. a new geographic knowledge base, called GeoNW which is obtained from the integration of GeoNames and WordNet data sources. GeoNW is used as a core resource for reaching most of the above objectives.

2. a new TR algorithm based on GeoNW and Natural Language Processing (NLP) techniques.

3. a new TD algorithm that uses GeoNW and information extracted from the context where the candidate toponym appears.

4. a new weighting model strategy for quantifying the relevance of geographic references over documents and queries that provides a simple mechanism of measuring the geographic similarity between a document and a query.

5. a new strategy for combining geographic and textual rankings in order to obtain a final ranked list of documents that fulfills both, textual and geographic restrictions. This final ranked list is supposed to better satisfy user needs.

6. a GIR framework, called GEIR (Geographic Enhanced Information Retrieval) which is developed as an extension of Terrier IR search engine. GEIR includes all the components that compose the GIR solution presented in this work. Its modular design allows to easily extend any of the components that conform a whole GIR solution providing a simple alternative for combining different approaches and comparing their results. GEIR can also be extended to analyze not only geographic features, but also other type of features like temporal or other specific domain features.

## 1.4 Organization of this thesis

In addition to this introductory chapter this document is divided in six more chapters. Chapter 2 discusses the related work in the literature. Chapter 3 describes in detail the proposed geographic knowledge base explaining its structure and how it is built as the integration of two well-known sources: GeoNames [108] and WordNet [76]. Chapter 4 presents the GIR solution here proposed. The toponym resolution algorithms along with the geographic relevance and ranking processes are explained in this chapter as components which compose a GIR solution. The design and implementation of the GEIR plugin and of the GeoNW knowledge base are explained in Chapter 5. Chapter 6 shows the results obtained during the experimentation step. Each independent component as well as the whole proposed GIR solution is evaluted by using different evaluation metrics and comparing with state-of-the-art approaches. The final chapter, Chapter 7, highlights the results of this work and discusses still open research lines.

# Chapter 2

# Related Work

In this chapter, we review previous and related work in the area of processing geographic information in unstructured texts. Section 2.1 defines the notions of geographic knowledge bases, specifically geographic ontologies and gazetteers and describes some examples of both kinds of resources. Most problems in GIR include the analysis of natural language which is the main focus of the Natural Language Processing (NLP) research area, thus Section 2.2 explains the application of NLP techniques in the resolution of those problems. Section 2.3 describes state-of-the-art techniques for recognizing and disambiguating toponyms in texts. These techniques are applied to the domain of GIR, namely the extraction of geographic information from documents and queries, a process that is described in Section 2.4. Section 2.5 discusses different strategies for comparing the geographic information present in texts in order to determine their similarity and consequently to provide a geographic ranked list of documents. This section also illustrates the application of geographic knowledge bases to the ranking process. Finally, Section 2.6 refers to standard techniques for evaluating GIR systems.

## 2.1 Geographic Knowledge Bases

Most GIR approaches are built on top of a knowledge base containing information about geographic references [11]. There are several works oriented to the design of such knowledge bases with different levels of complexity. For instance, there are geographical gazetteers consisting of plain descriptions, however these structures do not provide relationships among place names [53]. A more suitable kind of resource are the geographic ontologies which allow not only the main characteristics of a place associated to a toponym, but also the relationships between these toponyms [2, 37, 49, 50]. In this section we analyze in depth both, gazetteers and geographic ontologies as external resources for solving GIR problems.

### 2.1.1 Geographic Ontologies

The most widely accepted and common defnition of the geographic world is based on ideas of objects and fields [43]. To this end, the possible approach to overcome the problems of semantic heterogeneity and enhance semantic referencing of geographic information is the explanation of knowledge, by means of an ontology, which can be used for the identification and association of semantically corresponding concepts, because an ontology can explicitly and formally represent relationships between concepts and can support semantic reasoning according to different entities in the domain.

There are fundamentally different ways of thinking about the term "Ontology" in philosophy and information science. For example, Uschold and Gruninger in [103] define ontology as "a formal, explicit specification of a shared conceptualization", while according to Fonseca [36] "Ontologies are theories that use a specific vocabulary to describe entities, classes, properties, and functions related to a certain view of the world, they can be a simple taxonomy, a lexicon or a thesaurus, or even a fully axiomatized theory". In the same way, there are also definitions of the purpose of an ontology, for example (and probably the most accurate one in our scenario), the one given by [3]: "Ontologies serve to improve the accuracy of searching and enable the development of powerful applications that execute complicated queries, whose answers do not reside on a single web page."

The Semantic Web relies heavily on the formal ontologies that structure underlying data for the purpose of comprehensive and transportable machine understanding [3,58]. Therefore, the success of the Semantic Web depends strongly on the proliferation of ontologies, which requires fast and easy engineering and avoidance of a knowledge acquisition bottleneck.

In 2004, The World Wide Web Consortium released the Resource Description Framework (RDF) and the OWL Web Ontology Language (OWL) as W3C Recommendations [17, 55]. RDF is used to represent information, and to exchange knowledge in the Web. OWL is used to publish and share sets of terms called ontologies, supporting advanced Web search, software agents and knowledge management.

The searching and retrieval of geographic information is performed by the execution of geographic queries based on user input keywords. However, the execution of these keyword queries are not informative enough for GIR. Usually queries are very short making more difficult to infer their contained geographic information. Ontologies plays an important role to accomplish this task providing semantic description and referencing of geographic data. They contain, in a very structured way, domain knowledge and specific data regarding a certain subject field [103].

The basic design idea of ontology-based IR can be summarized as follows:

1. Establish the ontology of the related fields with the help of domain experts.

2. Collect data from the information sources with reference to the established ontology and store it in **prescribed format**.

3. In accordance with the ontology, the query converter transforms the search requests from user interface into **prescribed format**.

4. After processing, the result of the retrieval is returned to the user.

A Geographic Ontology is an ontology with spatial relationships among geographic features. A geo-ontology describes i) entities that can be assigned to locations on the surface of the Earth; ii) semantic relations among these entities that include, hypernymy, hyponymy, mereonomy and synonymy relation; and iii) spatial relations among entities (e.g., adjacency, spatial containment, proximity and connectedness).

Some geo-ontologies that can be used as a resource in GIR systems are:

- **Yahoo! GeoPlanet**[1]: is a resource for managing geo-permanent named places on earth. It provides developers the opportunity to make geographic aware applications, by the usage of unique geographic identifiers that allow through the Yahoo GeoPlanet web services to unambiguously tag geographic data across the web. Some of the information provided by Yahoo! GeoPlanet includes:

    - WOEID or Where-On-Earth IDentifiers: a set of identifiers, each one uniquely associated to a place on the Earth
    - Hierarchical containment of all places up to the Earth level
    - A set of known ZIP codes and its corresponding WOEID
    - Adjacencies: places neighbouring each WOEID
    - Aliases: synonyms for each WOEID

  Yahoo! GeoPlanet provides information for about six million named places globally. The coverage varies from country to country. It includes several unique administrative and historical areas; over two million unique settlements and suburbs, and millions of unique postal codes covering about 150 countries. The information is structured in a hierarchical way which allows to preserve the geographical containment relations. In this way, from a WOEID representing a place one can reach the full geographical context of a place.

- **Wikipedia: WikiProject Georeferencing**[2]: It is an active project, developed by the Wikipedia community, which aims to attach the geographic coordinates to the existing articles in Wikipedia that are related to geographic references (i.e., University of Bologna, New York City, Eiffel Tower, etc.). Currently, there are more than four millions of entries already associated with their corresponding geographic coordinates, from which more than one million belong to the English part of Wikipedia.

- **GeoWordNet** [42]: is a semantic resource built from the full integration of WordNet[3], GeoNames[4] and other semantic resources. GeoWordNet mainly combines the classes present in GeoNames and WordNet and then populates these with the corresponding entities. This dataset contains 3,698,238 entities, 3,698,237 part-of relations between entities, 390 concepts, 182 relations between concepts, 3,698,238 relations between instances and concepts, and 13,562 alternative entity names of which 10,042 are in English and the rest in Italian.

[1] https://developer.yahoo.com/geo/geoplanet/guide/index.html

[2] https://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en

[3] http://eprints.biblio.unitn.it/1777/

[4] http://www.geonames.org

- **Geo-WordNet** [21]: It is an automatic annotation of WordNet with geographical coordinates. Do not confuse Geo-WordNet with the above mentioned geographic ontology GeoWordNet (without -). In this case, the authors extracted the geographic synsets present in WordNet, with their holonyms and hypernyms and annotated them with their corresponding geographic coordinates. The geographic coordinates were obtained from the Wikipedia-World geographical database[5].

- **GeoNames**: is an open project for the creation of a world geographic database. It contains more than 10 million geographical names and consists of 8 million unique features. All features are categorized into one out of nine feature classes and further sub-categorized into one out of 645 feature codes. It integrates geographical data such as names of places in various languages, elevation, population and others, from various sources. Users may manually edit, correct and add new names using a user friendly wiki interface. This resource is probably one of the most used by the GIR and GIS communities when it comes to the necessity of a geographic knowledge base.

- **WordNet** [76]: It was developed at Princeton University as a complex lexical database of general English. Its last version (3.0) contains 155,327 words grouped in 117,597 synsets. A synset (set of synonyms) is a group of words that shared the same meaning. See Section 3.1.2. Among these synsets there are about 2000 that represents word locations, plus the meronyms of these that make around 7000 places in total. The main drawback from WordNet is the lack of geographic coordinates associated to place related synsets. Hoever, even when WordNet's main focus is not the geographic knowledge, it has been used by GIR community to address different problems. For example, Buscaldi [22] uses the WordNet ontology for applying query expansion to geographical terms, based on the synonymy and meronymy relationships. In [23], Buscaldi also showed that WordNet can be used for solving toponym ambiguity problems.

### 2.1.2 Gazetteers

Another GIR resource that also can be used for extracting and analyzing geographic information are the gazetteers[6]. It is an alphabetical list of place names with information that can be used to locate the areas that the names are associated with. There are three styles of gazetteers: alphabetical list, dictionary, and encyclopedic.

- The alphabetical list includes place names and locations as well as information typically found in atlases. For example:
  - **Place Name Sites**: It was created by the Association of British Counties. It contains more than 50,000 place names located in Great Britain.

[5] https://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Hauptseite/Wikipedia-World/en

[6] http://www.library.illinois.edu/max/collections/gazetteers.html

- Dictionary-style gazetteers include location information in the form of geographic coordinates or descriptions of spatial relationships with other places. For example:
  - **Canadian Geographical Names**: It was created by the Canadian Board on Geographic Names as part of Nataural Resources Canada. It contains place names and names of geographical features that appear on Canadian maps.
- Encyclopedic gazetteers will include all the information of a dictionary-style gazetteer but the information will be more detailed and may come in the form of articles written by area specialists.
  - **Getty Thesaurus of Geographic Names (TGN)**: It is a gazetteer of world place names primarily from the field of art history. On May, 2015 TGN contained more than two millions of place names.

### Discussion

In the present work we present a custom knowledge base (see section 3.2) built from a combination of the GeoNames and WordNet ontologies. The approach for the combination of these two ontologies is somehow inspired in the approaches used in construction of GeoWordNet and Geo-WordNet tools.

## 2.2 Natural Language Processing in GIR systems

Natural Language Processing (NLP) is a sub-area of artificial intelligence that deals with interpretation of human language [78]. At the beginning, the methods used in NLP had great acceptance and success, since application environments were more or less simpler. However, when applications were put into practice in uncontrolled environments with more generic vocabularies, many difficulties began to emerge. For example, the problems associated with polysemy and synonymy.

Geographic Information Retrieval requires converting common language terms to geographic concepts. Mapping informal geographic representations, such as place names of locations that people use as part of its natural language, to formal geographic footprint based on specific coordinates is not a trivial task, even with the help of gazetteers and geographic ontologies. While we rely on natural language to seek and retrieve geographic information, language is often not precise enough. Language has the ability to structure mental representations of space because it acts as a structural framework to describe physical space and locations. It is important to note this structural framework takes into account certain features but ignores others, thus reducing the total information provided. Some GIR systems are using NLP techniques for recovering, in some way, this information that it is not explicitly provided. In [70], Manning describes the different tools that compose the Stanford CoreNLP toolkit, which can be used for geographic analysis in unstructured text. Also in [15, Chapter 7] Bird explains the use of NLP techniques for extracting information in texts.

Some aspects of conventional IR technologies need to be improved or even replaced by new more suitable ones, so that modern web engines could better understand the spatial semantics within both web documents and user queries. In the process of understanding this spatial semantics, Natural Language Processing (NLP) plays a main role. The recognition and grounding of place names provide important information that is potentially unavailable to a search system based purely on text. Without NLP components the solutions to problems like synonym detection, ambiguity resolution and accurate toponym expansion would be considerably harder. Because of this, most GIR systems use NLP techniques in pre-processing stages, at least to some extent.

In [99], Strzalkowski established a general architecture of an NLP-IR system, in which an advanced NLP module is inserted between the textual input (documents and queries) and the database search engine. The proposal consists in processing text with a sequence of a part-of-speech tagger, a lexicon-based morphological stemmer, and a fast syntactic parser:

- **Part-of-Speech (POS) Tagger**: POS tagger allows the resolution of lexical ambiguities in a running text, assuming a known general type of text (e.g., newspaper, technical documentation, medical diagnosis, etc.) and a context in which a word is used. This, in turn, leads to a more accurate lexical normalization or stemming. It also is a basis for a phrase boundary detection.

- **Lexicon-based normalization**: Word stemming has been an effective way of improving document recall since it reduces words to their common morphological root, allowing more successful matches. On the other hand, stemming tends to decrease retrieval precision. Special treatment is needed to prevent situations where otherwise unrelated words are reduced to the same stem.

- **Syntactic Parser**: Parsing reveals finer syntactic relationships among words and phrases in a sentence, relationships that are hard to determine accurately without a comprehensive grammar. Some of these relationships do convey semantic dependencies.

Certain types of phrases are extracted from the parser trees and used as compound indexing terms in addition to single-word terms. The user natural language request is also parsed, and all indexed terms occurring in it are identified. One important remark is that removing low-quality terms from the queries is at least as important as adding synonyms and specializations. After the final query is constructed, the database search follows, and a ranked list of documents is returned. It should be noted that all processing steps, those performed by the standard system and by the NLP components, are fully automated, and no human intervention or manual encoding should be required.

### Discussion

The importance of NLP techniques in our solution is paramount. NLP plays a key role in the toponym recognition (see 4.1.1), toponym disambiguation (see 4.1.2) and toponym weighting model (see 4.2.1).

## 2.3 Toponym Resolution

Finding the geographic references in an unstructured text is called toponym resolution. The process is divided in two main sub-processes: i) recognizing the geographic locations in the text (toponyms), which is known as Toponym Recognition (TR) and ii) disambiguating the recognized toponyms, which is knowm as Toponym Disambiguation (TD).

The main reason toponym resolution is a challenging task are the several ambiguities in natural language. According to Amitay [5], ambiguities can be classified in two main groups: i) non-geo/geo; and ii) geo/geo. The former corresponds to places names that come from natural language (i.e., *Nice* in *France* or nice, the English adjective) and it is mostly solved during TR algorithm. The latter, tackled by the TD algorithm, refers to different places around the world with the same name (i.e., *Havana, Cuba* or *Havana, Texas*).

This section reviews state-of-the-art approaches related to both, the TR and the TD sub-processes.

### 2.3.1 Toponym Recognition

Toponym Recognition (TR), also known as geo-parsing, geo-tagging or toponym extraction is the process of analyzing unstructured texts, in order to find geographic references in it.

The core difficulty associated with the TR process consists of the different kinds of ambiguity associated with natural languages. For example cases where it is not possible to determine if a specific name is related to a geographic term or to another kind of reference, such a person name. This kind of ambiguity, which was defined by Amitay [5] as geo/non-geo ambigutity, can lead to undesired results, especially in geographic search engines, due to the fact that these systems rely on geographic references found in documents to satisfy the user geographical query. Therefore, removing geo/non-geo ambiguities is crucial for a successful TR. Note that the amount of geo/non-geo ambiguities is affected by the level of granularity of toponyms considered in the procedure. A TR process for country-level toponyms could be considered easier than one for city-level toponyms, due to the comparatively smaller number of country toponyms, which provides fewer opportunities for a geo/non-geo ambiguity.

Many different types of entities can be considered geographic references. Perhaps the most obvious are geopolitical entities, such as countries (e.g. *Italy*) and administrative divisions (e.g., *California, Florida*), as well as populated places such as cities and towns (e.g., *Palo Alto, Miami*). Other types of region locations can include postal codes and municipal areas. At a smaller scale, various hyperlocal locations could be considered, such as streets (e.g., *Via Zamboni*), street addresses (*Via Mura Anteo Zamboni 7*), street intersections (e.g., *62nd street and 79 Ave*), city centers (e.g., *downtown Miami*), and buildings (e.g., *Empire State building*). In some applications, natural geographic features would be locations of interest, such as parks (e.g., *Dolomiti Park*), rivers (e.g., *Nile*), and mountains (e.g., *The Alps*). Finally, there are references to imprecise areas (e.g., *east coast, southern France*). Each of these location types affords different kinds of contexts that enable readers to understand that a location is being referred to.

### Knowledge-based methods

The text is parsed literally, and searched for matching occurrences of a predefined known set of toponyms. This known set usually comes from a gazetteer or a geographic ontology [48, 49, 105]. A key aspect in this kind of methods is the quality of the data sources. The incomplete and noisy nature of the data can lead to false positives and false negatives. Also place names and administrative boundaries are constantly changing, thus there is a need of managing an update process of the knowledge base.

One of the most known approaches in this group is Amitay's *Web-a-Where* method [5]. In this approach the author defined a custom gazetteer based on the integration of different sources. In addition, a set of rules are defined in order to overcome classic problems of this kind of method, as the presence of places with names that are common English words like "To" Myanmar), or "Of" (Turkey). Another interesting feature of Amitay's approach is the relation between place name occurrences and the population of the location, one of the examples given by the author is the word "Atlantic" (Iowa), which was not considered as a place name, although it appears commonly in the analysed collection (10,976 occurrences) its population falls below that number (7,474).

A similar approach is proposed by Katz [52]. However, in contrast to several other approaches which focus a lot on the development of preprocessing techniques for correctly detecting the geographic references in the text, they use very basic mechanisms for TR process. In this approach, all capitalized words/phrases are detected and then using a pre-generated case dictionary, candidates with a very low probability of being toponyms are removed. The dictionary consists of a list of tokens with the occurrence frequencies as uppercase andlowercase variant. If the candidate is more often written in lowercase it is removed from the set of candidate toponyms that will be analyzed in the TD process.

### Natural language processing

In an attempt to gather a better analysis of the context and increase the effectiveness of the recognition tasks, some approaches make use of Natural Language Processing (NLP) tools (e.g., Named-Entity Recognition (NER) and Part-Of-Speech (POS) tagging) [35, 70, 102]. These approaches can be roughly classified as either rule-based [23, 31, 34, 86, 115], statistical-based [64, 101] or machine learning-based [52, 62]. Many of these ideas are also used during the disambiguation part. In the next section, we describe with more detail some of these solutions.

## 2.3.2 Toponym Disambiguation

Toponym Disambiguation (TD), also known as geo-coding, can be defined as the task of associating an ambiguous toponym with the (unique) geographic location that best matches in a given context. In general, TD approaches share the common idea of giving a score to each possible location alternatives and finally selecting the one with the highest score. Depending on the method used to calculate the score, TD methods may be grouped into three main categories (see [19]):

- **map-based**: methods that use an explicit representation of toponyms on a map, for instance to calculate the average distance of unambiguous context toponyms from referents;
- **knowledge-based**: methods that exploit external knowledge sources such as gazetteers, Wikipedia or ontologies to find disambiguation clues;
- **data-driven or supervised**: methods based on machine learning techniques.

### Map-based methods

Map-based methods consider mainly the coordinates of the places appearing in context. These methods are usually very sensitive to changes in context; therefore, it is necessary to remove places that are very far on average from the others in the context [95], or to include external knowledge, such as the location of the source of a text [20]: for instance, if the toponym *London* is found in a *Ontario, Canada* based newspaper, it is more likely that it is referring to *London, Ontario* rather than *London, UK*. The source of the information, has been proved to be of higher importance in this disambiguation approach when applied to locally restricted text collections: in [20], it has been shown that 76.2% of the places mentioned in an Italian newspaper are located within 400 km of the city where the newspaper is published.

The main idea behind the map-based methods is based on the use geo-tagged documents to associate toponym occurrences with a particular a map grid cell containing the document location. In this methods the strategy for the map division in cells proved to be key in order to obtain accurate results [62, 64, 101].

### Knowledge-based methods

Knowledge-based methods use toponym metadata information to pick the most likely place. In the case of the work by Amitay [5] a sets of heuristic built around the place population is defined in order to properly make the disambiguation. The selection of the population metadata feature is used, is based on the empirical assumption that more populated places are more likely to be mentioned.

Another knowledge-based approach uses the hierarchical administrative structure of the places occurring in the context. The underlying idea is that the places in the context tend to be grouped in common regions or geographical areas. Some examples of hierarchy-based algorithms are [13, 23].

Aligned with this idea Katz and Shill [52] proposed an disambiguation strategy relying on a set of heuristics. These were based on a set of rules that use the information obtained from a custom gazetteer created from the GeoNames database.

### Data-driven methods

Data-driven methods are not commonly used in TD, mainly because of the lack of good and large training datasets, and consequently, the inability to classify unseen toponyms. However, this approach has the huge advantage of exploiting non-geographical content, such as in the work of [87], where events are used to build a probabilistic model, using the spatial relationships between non-geographical entities and places; for instance, if some known person or organization is based in a place, their presence in the context of the toponym may represent an important clue (for instance, *political activities* in the context of *Washington* may suggest that it is *Washington D.C.* rather than *Washington State*).

Also Katz and Shill [52] proposed a data-driven approach based on a classifier trained as part of a machine learning process. The authors noticed that adding more rules to further improve their heuristic approach (described above) will increase the risk of over-fitting the algorithm. Thus, this machine learning based approach relies on a number of features and a classifier which is trained using manually annotated and disambiguated training documents. The test collection as well as the results obtained from their proposals have been used in this work during the evaluation process.

Interesting approaches are also those based on language models. For example in the work of Wing and Baldridge [110]. In this work the authors proposed a enhancement of a map-based strategy, the underlying idea is to have a trained algorithm that associates cells from a map grid with each document. The training process was made using geolocated Wikipedia contents, while for the map cells association the authors proposed several probabilistic models that considered the co-occurrences of toponyms with other (possibly non-geographic) words. Later on, an extension of this work by Roller and Wing [89] demonstrated that some enhancements were possible by using an unevenly distributed map grid. The main intuition behind this improvement is that usually place references are not distributed equally around the globe.

### Discussion

In the current solution we present a knowledge based approach for the TR process (see 4.1.1). Such approach is built on top of GeoNW which acts as the principal source of toponym elements, however in order to enhance the precision of the recognition process we define a set of heuristic rules based on NLP techniques that allow removing ambiguities specially of the *geo/non-geo* kind. The toponym disambiguation process (see 4.1.2) uses a technique very similar to the heuristic-based one proposed by Katz [52]

## 2.4 Geographic Information Extraction

Another definition of Geographic Information Retrieval was proposed by Jones and Purves [51]. They defined GIR as the "provision of facilities to retrieve and rank by relevance documents or other resources from an unstructured or partially structured collection, on the basis of queries specifying both theme and geographic scope". This definition carries out some challenges that must be addressed by the Geographic Information Retrieval community. These include:

- identification of geographic references (toponyms) in documents and associating these terms with appropriate geographic locations

- indexing large collections efficiently for searching on both thematic and geographic content

- development of search engines and algorithms which can exploit such indexing systems

- techniques to properly combine geographic and thematic relevance

As mentioned before, there are four basic features relevant for GIR systems: place-names (e.g., *Italy, Miami, Toronto, etc.*), geographical relations (e.g., *south, north, near, far, etc.*), geographical concepts (e.g., *lakes, cities, mountains, etc.*) and geographical adjectives (e.g., *Italian, Canadian, northern, etc.*). These geographical terms participate in the definition of the geographical footprint of documents and queries. However there are other terms that are not considered explicitly geographic terms but contain a geographic sense, such as, names of people or names of organizations. Therefore, documents and queries which do not mention geographical terms explicitly may still have a geographical footprint.

### 2.4.1 Document Processing: Geographic Footprint

In GIR applications the geographic footprint is typically the only quantitative geographic characteristic for describing the geographic information in unstructured texts. It is usually a geometric representation expressed in geographic coordinates (latitude, longitude). Examples of geographic footprint representations are:

- **Points**: it keeps a general sense of location but not extent or shape.

- **Polygons**: it identifies location, extent and shape varying the degree of precision. The minimum bounding rectangle (MBR) is the most common polygonal geographic representation in GIR systems [85].

Assigning the geographic footprint to documents includes the correct identification of the toponyms. This process is known as toponym resolution and it aims to solve two well-known problems: Toponym Recognition (TR) and Toponym Disambiguation (TD) (see Section 2.3).

The geographical footprint of a document can help in retrieving documents by imposing geographical restrictions on the search query. One of the first works in this subject was proposed by Woodruff and Plaunt [111]. They compute the geographic footprint of a document based on the references that appear in the text. Their method is based on disambiguating the toponyms into their respective bounding polygons. The geographic footprint of the document is then computed using the overlapping area of all the polygons, trying to find the most specific place that is related to all the place references mentioned in the text.

Ding [34] also proposes a technique for determining the geographic footprint. In this case, he defined the geographical footprint of a web resource $\omega$ with the help of a hierarchical gazetteer. The proposed approach introduces two concepts: Power and Spread to decide the geographical footprint of web documents. They proposed two kinds of resources for the calculation of a web page geographic footprint: the locations of the pages linking to the web resource $\omega$ and the place names appearing in $\omega$.

Amitay proposed a system named Web-a-Where [5]. It uses a hierarchical gazetteer as their knowledge base too. Firstly they generate a hierarchical relationship for every toponym appearing in the document as $A/B/C$ (e.g. New York/USA/North America). Assigning to each node its value of importance, they sort the nodes by these values and then select the most relevant toponym as the geographical focus of the document.

Martins and Silva [72] uses the PageRank algorithm to infer a single global geographic footprint for each document. First, geographical references are extracted from the text and associated with the corresponding concepts in a geographical ontology. Every document is represented by a set of features corresponding to toponyms from the text. Each feature associates a weight to a set of concepts in the ontology according to their occurrence frequency. A graph is used to represent the association between the geographical references and the ontology concepts in order to apply the PageRank algorithm for determining the geographic footprint. Each concept is represented as a node in the graph, and a relationship statement is represented by two directed edges. Different types of relationship in the ontology correspond to different edge weights in the graph. A PageRank formula based on edge and node weights is used to compute the ranking score for each node in the graph. The toponym with the highest weight is finally assigned as the geographical scope of the page.

Pouliquen [85] and Leidner [60] defined a document geographic footprint at country level. They employed the country of the document publication and the countries of the most important unambiguous toponyms extracted from the text.

In 2009, Campelo and Baptista [26] proposed a model for detecting toponyms in web documents, based on a set of heuristics. In the proposed approach they introduced the concepts of confidence factor and confidence modifier, which aim to measure the probability of a detected geographic reference of being a valid reference and being associated to a correct place, even when there exists ambiguity.

Many factors can be used to decide the geographic footprint of a document such as textual information in the document, hyperlinks, web environment, geographical references and so on. Automatic assignation of geographical footprint to documents remains a complex problem in GIR that is attracting more and more attention from researchers. In [6] we can find a comparison of different approaches related to this thematic.

### 2.4.2 Query Processing

While an effective document processing technique allows to extract accurate information from documents in order to find those that are relevant to the user, effective query processing method is a key aspect for interpreting what the user is looking for.

The type of query in a standard IR search engine is based usually on natural language, in contrast to the more formal approach in Geographic Information Science (GIS), where specific geo-referenced objects are retrieved from a structured database. In a GIR system, a geographic query is also expressed in natural language and it includes a theme (**what** to search), a geographical location (**where** to search) and the spatial relationship (what is the relationship between **what** and **where**). The spatial relationship indicates how the object of interest is related to the specified geographical location. Common spatial relationships are distance, directional and topological relationships. According to this, [48] defines a geographical query as a tuple <**what, relation, where**>.

Several authors have studied what users are looking for when submitting geographic queries [38]. One of the main conclusions of these studies is that the structure of geographic queries consists of thematic and geographical parts, with the geo-part occasionally containing spatial or directional terms.

From a geographical point of view, Kohler [91] provides a research about geo-reformulation of queries, concluding that the addition of more geographic terms in the query is commonly used to differentiate between places that share the same name. This is also known as query expansion using geographic entities. The purpose of query expansion is to make the user query resemble more closely the documents it is expected to retrieve.

Currently, query expansion usually refers to content and typically is limited to adding, deleting or re-weighting of terms. For example, content terms, from documents considered relevant, are added to the query and the weights of all terms are adjusted in order to reflect the relevance information. Thus, terms occurring predominantly in relevant documents will have their weights increased, while those occurring mostly in non-relevant documents will have their weights decreased. This process can be performed automatically using a relevance feedback method.

In the literature, there are some works that have addressed the spatial query expansion. Cardoso [27] presents an approach for geographical query expansion based on the use of feature types, readjusting the expansion strategy according to the semantics of the query. In [22], Buscaldi uses WordNet during the indexing phase by adding the synonyms and the holonyms of the encountered geographical entities to each documents index terms, proving that such method is effective. While Stokes concludes that significant gains in GIR will only be made if all query concepts (not just geospatial ones) are expanded [98].

Once the query is processed its geographic footprint representation is defined according to the one used for documents in order to lay the groundwork for the subsequent similarity analysis between a query and a document according to their geographic information.

### 2.4.3 Indexing methods

Generally GIR solutions rely on hybrid indexing process, in which separated processes tackle the thematic indexing and the geographic indexing [44, 114].

In the particular case of the geographic indexing, the most common approach is the use of a spatial index structure like R-Trees or PK-Trees [107]. The use of such structures has numerous advantages, being one of the most relevant the possibility of linking documents with place hierarchies. On the other hand the main drawback of the use of such structures resides in the performance impact. There have been several attempts aiming to the design of improvements of these structures [61, 88].

#### Discussion

In the provided solution no spatial indexing structure is used. Instead, the geographic indexing makes use of a classic direct index and inverted index structures. The main reason is that the current solution inherits the index structures available in the Terrier framework. Although this is not the common approach, there are some advantages related to this decision, for example, (i) the performance of the inverted index structure for finding documents related to a particular term and (2) the re-usability of the indexing mechanisms from the Terrier framework, for example, the Hadoop based indexing, useful for large document collections. Indexing as in standard IR gives also the advantage of combining similarly computed scores when it comes to combine textual and geographic relevance (see 4.4).

There is, however, one major downside of using classic IR indexing for geographic terms: the lack of a administrative hierarchy representations makes difficult to identify, for example, that documents mentioning *Toronto* are relevant for queries containing the term *Canada*. To overcome this issue our technique reports, during the document footprint extraction, not only the terms actually mentioned in the documents, but the ones implicitly mentioned as well (see 4.2).

## 2.5 Relevance Ranking

Another essential part of IR is assigning a relevance score to documents to represent how well they fulfill users information needs. This is done by attempting to emulate how a user would judge a document relevant to a query. When calculating geographic relevance the motivation is the same, but the methods employed are quite different. There are a number of approaches to query-document geographic relevance, however there is not currently a consensus about what is the best alternative to apply. GIR systems can either use one or a combination of them.

According to Overell [81], methods for computing the geographic relevance of documents with respect to a query can be divided in three main groups: footprint methods, distance methods and topological methods.

### 2.5.1 Footprint methods

These methods are directly based on the geographic footprint that represents the geographic information in documents and queries.

**Footprint based on polygons.** Beard and Sharma [12] compute the geographic relevance as the area of overlaps between the polygons that represent the query and the document. They consider three possible spatial relations for ranking purposes (Figure 2.1):

1. the document footprint **is inside** the query footprint
2. the document footprint **overlaps** the query footprint
3. the document footprint **contains** the query footprint

In the first case, where document footprint is inside the query footprint the spatial score is assigned on the basis of the ratio of both areas:

$$\rho = \frac{document\_footprint(area)}{query\_footprint(area)}$$

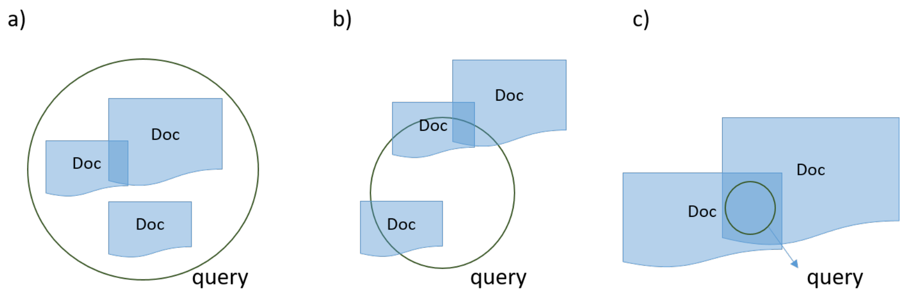

Figure 2.1: Possible spatial relations among documents and query. a) query footprint contains document footprint, b) query and document footprints overlap, c) document footprint contains query footprint

A ratio of 1 represents the unlikely an exact match. A score close to zero means the document covers only a small area of the query, which means the document refers only some of the geographic references present in the query.

In the second case, the rank is based on two criteria:

- percentage of overlap between the document and query, with reference to query area, called $q$ and
- percentage of non-overlap between document and query, with reference to document area, called $r$

Then, the score is computed as:

$$\rho = \frac{q}{r + 100}$$

Again a value of 1 is a perfect match. The greater the overlap region, the higher will be the rank. However, this expression penalize the documents with large non-overlapping region, with the intuition that these documents are more related to other locations than to those include in the query.

Finally, the third case is similar to the first one with rank assigned on the basis of the ratio of areas between the query and the document.

$$\rho = \frac{query\_footprint(area)}{document\_footprint(area)}$$

Again a score of 1 is a perfect match and scores close to 1 indicate a document refers mainly to those places referenced in the query. In cases where document regions are much larger than query regions the document receives a low rank. Intuitively, this means that the document is not very specific with respect to the query.

An important remark is that this scheme, gives no importance to documents having a footprints adjacent to (are close to) the query footprint. In some scenarios this may be a drawback.

This approach may be hard to apply in unstructured data where calculating geometric representations of the documents and query footprints is challenging. However, it can be used as a base for further approaches.

**Footprint based on representative points.** Another alternative in this group of methods was proposed by Fu and Jones [37]. They compute the footprint similarity based on the distance between representative points for the places involved. If the place is represented by a polygon, the centroid is used as the representative point and the origin is used in the case of polylines, although in general, approaches based on this technique rely on the place geographic coordinates. Accordingly, they have the following criteria to measure the similarity of two footprints $F_1$ and $F_2$.

$$\sigma(F_1, F_2) = \begin{cases} 1.0 & if\ F_1 = F_2 \\ min(1.0, \frac{tol}{dis}) & otherwise \end{cases}$$

If $F_1$ and $F_2$ are the same, they have a perfect match equal to 1. Otherwise, a tolerance $tol$ is used along with the distance $dis$ between $F_1$ and $F_2$ to determine the matching score. Since $tol$ is a constant, the bigger $dis$, the lower is $\sigma(F_1, F_2)$. If $dis < tol$, $\sigma(F_1, F_2) = 1$.

Experiments based on this technique were developed using TGN[7] and SABE [8] datasets [37]. The similarity measure was used for comparing the geographical elements from both datasets in order to recognize equivalent terms. Taking as a initial value of $tol = 10km$, the 80% out of missing matching pairs were classified as correct and the remaining 20% were classified as uncertain. This indicates that the $tol = 10km$ is too rigid and needs to be relaxed. In next experiments they take $tol = 15km$ and $tol = 20km$ improving the final results. Authors suggested a deeper research finding the optimal value of $tol$ parameter.

Several GIR systems represent footprints as MBRs and keeping all rectangles aligned to a grid. This has many advantages including that the footprints can be represented with two points (opposite corners) rather than four points (every corner) and the footprints can be stored and searched more efficiently in a spatially index. The main disadvantage with this scheme is that dependencies related to the orientation of locations, the size, shape and relationship between MBRs can vary significantly. Figure 2.2 is an example of this problem. Black points are the set of points corresponding to each country. Using these points two rectangles are built, one for representing Portugal and another for representing Spain. Finally, analyzing the relation between both regions we obtain that almost all the Portuguese cities are within the Spain region, which is a mistake. More over, we notice that in the representation of Spain the Canary Islands have been left out, having included these in the MBR representation of Spain would have caused an even bigger imprecision.

Considering this problem, a derived conclusion is that is not recommendable to use the MBR scheme as the only way of compare geographic terms. Andrade and Silva [7] proposed a geographic similarity measure that combines the MBR representation with the information obtained from a geo-ontology. Given a query scope $S_q$ and a document scope $S_d$ they compute the similarity between two geographic scopes based on three main concepts:

- **Inclusion:** For testing if $S_d$ is inside $S_q$ and weighting the relationship

[7] It is a structured vocabulary containing names and other information about places. See http://www.getty.edu/research/tools/vocabularies/tgn/about.html for futher information

[8] It is the predecessor of the current project EuroBoundaryMap which provides a European geographic database for administrative and statistical regions. Further information can find in http://www.eurogeographics.org/products-and-services/euroboundarymap

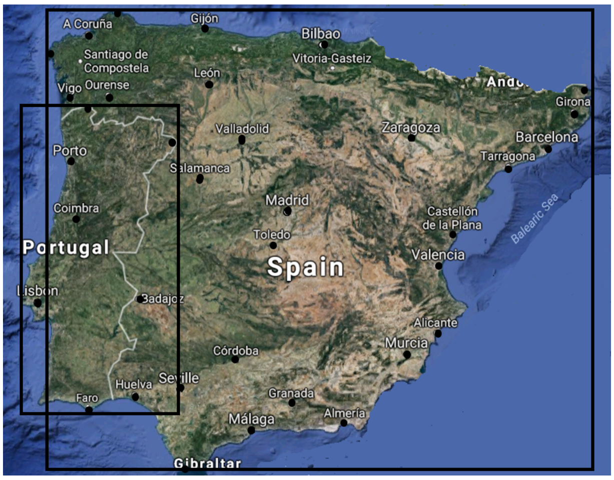

Figure 2.2: Footprints of Portugal and Spain using MBRs

degree between both scopes by the number of descendants in the geo-ontology:

$$Inclusion(S_q, S_d) = \begin{cases} \frac{NumDescendants(S_d)+1}{NumDescendants(S_q)+1} & if\ S_d \in S_q \\ 0 & otherwise \end{cases}$$

This formula returns values in $[0;1]$, with the maximum value reached when both scopes are equal and the minimum when $S_q$ is not an ancestor of $S_q$.

- **Proximity:** It is the inverse distance:

$$Proximity(S_q, S_d) = \frac{1}{1 + \frac{Distance(S_q,S_d)}{Diagonal(S_q)}}$$

where the Euclidean distance is normalized by the diagonal of the MBR of the query scope.

- **Siblings:** It is a binary function that tests if $S_q$ and $S_d$ are siblings in the geo-ontology graph:

$$Siblings(S_q, S_d) = \begin{cases} 1 & if\ \exists S_x : parent(S_q) = S_x \wedge parent(S_d) = S_x \\ 0 & otherwise \end{cases}$$

The final geographic similarity function is defined as the linear combination:

$$GeoSim(S_q, S_d) = \beta * [Inclusion(S_q, S_d) + Proximity(S_q, S_d)] + (1 - \beta) * Siblings(S_q, S_d)$$

where $0 \leq \beta \leq 1$ so that the final value lies in the interval $[0, 1]$.

### 2.5.2 Distance methods

If the geographic footprint of a document/query is defined as a collection of points, the relevance between the query and each point in a document can be measured using some distance definition. In general this process is based on calculating the distance from all points in the query to all points in the document and combining them into a final relevance judgment. A common way of performing this combination is to simply take the location (point in the document) with the greatest relevance score.

For example, Hauff [45] proposed to add geographical knowledge in order to improve a traditional IR system. Their approach consists on:

1. To carry out document retrieval to find **content relevant** documents. For example, for the topic *Ford dealer near Florida*, this step should return a ranked list of documents discussing *ford dealer*, not necessarily near Florida.

2. To filter this ranked list based on geographical relevance. For each content relevant document, determine if it is also geographically relevant. If it is not, then the document is removed from the list.

Each query is analyzed for extracting locations and if a location corresponds to a country/city, its boundaries are applied as location coordinate restrictions. The locations found in each document are matched against these coordinate restrictions. A document is removed from the result list, if it does not contain any locations within an appropriate distance from the boundaries.

According to this approach the ranked list depends on the content analysis and the geographic information is used to reduce this list. This technique was evaluated on GeoCLEF 2006 [40] and its results did not improve the baseline algorithm (using the content only analysis). The author made a manual evaluation of the relevant documents of the first eight GeoCLEF 2006 topics which revealed that the exact location phrases mentioned in the title query also occur in almost all relevant documents. This makes a geographically enhanced approach unnecessary and also explains the similar results between the baseline and the geographically filtered results for the title queries.

GIR systems vary in their distance calculations, some systems project the Earth's surface onto a two dimensional plane and apply Euclidean Geometry to calculate distances. For example, Park [84] defined two functions for measuring the distance between documents and query: geographical size of a set of spatial objects and geographical distance between the query and this geographical size.

The geographical size $GSize(S_k)$ is defined as:

$$GSize(S_k) = \max_{(o_i,o_j)\in S_k\dot{S}_k, i\neq j} minDist(o_i.loc, o_j.loc)$$

where $S_k$ is the set of geographical objects in a document, $o_i$ is a geographical object in $S_k$ with its corresponding location and $minDist(A, B)$ is the minimum Euclidean distance between $A$ and $B$.

The geographical distance between the query location $q.loc$ and $S_k$ is defined as:

$$GDist(q.loc, S_k) = \min_{o_i\in S_k} minDist(q.loc, oi.loc)$$

Finally the score of a relevant document with respect to a query is determined as:

$$Score(q,d) = \frac{1}{1 + ln(1 + GDist(q.loc, S_k) * GSize(q.loc, S_k))}$$

where $S_k$ is the set of geographical objects retrieved from the document $d$. Because this approach is one of the most recent included in this kind of methods, there is no a trusty evaluation of the strategy for proving its effectiveness.

According to Montello [77] a drawback of using absolute distance as a method for calculating relevance, is that the human appreciation of distance is relative. For example, the appreciation of the distance between Madrid and Bologna varies for a user situated in New York or in Bologna.

### 2.5.3 Topological methods

As stated before, the human appreciation of the relationship between locations is asymmetric and inconsistent, and as such does not easily map into a metric space [77]. When geographic relationships are used in queries they are often ambiguous terms such as near or close. There are many ways of estimating these measures: for example the travel times between locations (in minutes and hours) or the required method of transport (walking, driving or flying).

Another method of modelling the relationships between locations is looking at topological distance. Topology, in this case, refers to how locations are connected. We can distinguish two kind of topologies: physical and political topology. Both of them differ and are components in judging relevance. For example, the British Isles are physically composed of two islands: Great Britain and Ireland; politically, the United Kingdom consists of four countries: Scotland, England, Wales and Northern Ireland, while the Republic of Ireland is a separate nation. These overlapping interpretations of topology add great complexity to the modelling of geographic relevance.

Also, we can distinguish the vertical topology which represents the political hierarchy of locations. From a geo-political point of view, a country is composed by regions or states, which are composed by provinces, and provinces are composed by cities and so on. This hierarchy can be represented by a tree where the root represents the whole earth, and the leaves represent places without further political divisions. In this hierarchy one can define vertical and horizontal topological distances. The vertical distance is the one established between elements in the same path to the root level, while horizontal distance is the one established between element that share a common ancestor but are in separated paths.

Some metrics for measuring the distance between locations according to both representations (vertical and horizontal topology) have been proposed. Martins [93] assigns normalised values between 0 and 1 to the vertical topology similarity, adjacency (based on horizontal topology), containment (based on vertical topology) and Euclidean distance. The geographic similarity is then defined to be the convex combination of these values. Rodriguez and Egenhofer [8] represent classes of geographic object in a hierarchy. They define the Matching-Distance that can be applied to a geographic ontology to measure how similar geographic classes are.

### 2.5.4 Geo-ontologies based methods

Alani and Jones [49] propose a technique based on those aspects of geographical space that are concerned primarily with location and proximity. They focused on the use of geographical hierarchies in combination with Euclidean distances. Search expansion methods based on these aspects of space will automatically find connected places and will tend to give those higher priority than to the disconnected places. Euclidean distance whether in map-grid space or as measured on the Earth's surface leads to a ranking based on physical proximity, and introduces the possibility of constraining the expansion of a search for similar places according to specified distance thresholds.

While Euclidean distance is probably the simplest and most used way of measuring locational similarity, it fails to consider some physical and political factors. For example, in large countries, the distance between an element and its ancestor may be large, even though both places are related. Likewise, in a small country two separated cities may be very close and be related only by the fact of belonging to the same country. Moreover, the fact of designing a single hierarchy relating a group of places is also non trivial. The most widely used hierarchy is the geo-political one. However, there can be defined other types of hierarchies like the geo-physical ones, for example, *Sicily* may be a descendant of *Italy* or of *The Mediterranean Sea.*

The geo-ontology based methods are based on non-common super-classes of geographical poly-hierarchies using generic $part - of$ relations that may be interpreted spatially as inside or overlap.

Thus, the Hierarchical Distance Measure (HD) between query place $q$ and candidate document place $c$ is defined as:

$$HD(q,c) = \sum_{x \in q.PartOf - c.PartOf} \frac{\alpha}{L_x} + \sum_{y \in c.PartOf - q.PartOf} \frac{\beta}{L_y} + \sum_{z \in q.PartOf - c.PartOf} \frac{\gamma}{L_z}$$

The $L_x$, $L_y$ and $L_z$ values represent the hierarchical levels of the individual places within their respective hierarchies. The set of places $x$ are those distinctive super-parts of the query term that belong to it but not to the candidate, while places $y$ are the distinctive super-parts of the candidate that are not shared with the query. The places $z$ are the query and candidate terms themselves. The sets of terms $q.PartOf$ and $c.PartOf$ refer to the transitive closure of the super-parts of $q$ and $c$ respectively in the part-of hierarchy. The weights $\alpha$, $\beta$ and $\gamma$ provide control over the application of the measure. In particular the weights $\alpha$ and $\beta$ provide the option of asymmetry. This same notion of similarity has been later extended by Lv [66] by including the depth of each element in the ontology. In general when applying the hierarchical distance measure, the distance between a query and a document increases according to the number of non-common parents, e.g., the distinguishing regions. The intuition followed by this approach is that the difference between elements in the top of the hierarchies should be more noticeable in comparison with the difference in the lower levels. It should be noted that the formula measures distance explicitly with regard to distinguishing super-parts, while closeness is regarded as implicit within the branching structure of the hierarchies.

In order to keep a balance between the Hierarchical Distance and the actual physical distance. The TD values can be combined with values resulting from the Euclidean Distance measure (ED).

The two locational distance measures can be combined in a weighted combination as the Total Spatial Distance (TSD):

$$TSD(q, c) = w_e ED(q, c) + w_h HD(q, c)$$

where $w_e$ and $w_h$ are weights of the ED and HD respectively. These weights lie in the range 0 to 1. In order to calculate a weighted combination of the individual distance measures as above, it is necessary to normalise both of the measures to a range between 0 and 1 prior to use.

### 2.5.5 Combination of thematic and geographic rankings

As stated before, the final results in GIR system are calculated after the combination of thematic and geographic rankings. There have been several approaches for tackling this task. Probably a very intuitive one is the approach of Martins et al. [73] that proposed a weighted sum of the thematic and geographic scores of each document in the results. Such approach was inspired in the seminal work of Shaw et. al. [92] that introduced the well known *CombMNZ* technique for the combination of relevancy rankings. Also, Martins et al. studied various alternative combination functions like the product or the maximum of the two individual scores [72, 73].

Another much older technique is the *Borda method* [33] invented back in 1791 for the combination of the vote counts obtained by candidates in different ballotages. A comparation of this technique with the previously mentioned *CombMNZ* is presented by Palacio et. al. in [82] as part of an evaluation approach for a complete GIR system.

Another approach was proposed by Yu and Cai [113] that aimed for a design of a query based combination method. In this case, the idea is to include in the combination characteristics of the query like the size of the geographical area covered. This approach demonstrated to have a positive impact in the final results, although it ultimately relies as well on a weighted relevance combination.

However, as pointed by Palacio et al. [83] one of the main drawback of pure arithmetical combination techniques is that the values computed for thematic and geographical scores tend to distribute in different ways, even after standard normalization. Their main stated that the merging of scores obtained make sense only when the relevancy calculation formula is supported by similar methods. To overcome this problem, in [83] the authors proposed a score normalization technique based on the term frequency distribution by spatial regions.

A novel approach was introduced by Van Kreveld et al. [104] with the idea of using distributed ranking methods. The intuition behind this theory is that when combining two or more ranking values, depending on the combination function, elements with similar values may be actually quite dissimilar. For example, a document very close to the query from the thematic point of view and another document very close to the query but from the geographic point view, might end up with similar ranking values during the combination process. To avoid this issue Van Kreveld's approach calculates the euclidean distance between the query and the documents to find the nearest one, which will become the first element in the ranking. From this point on, the rest of the documents will be added to the ranking according not only to the distance to the query but to the already ranked elements. Van Kreveld's contribution also included other alternatives to the main algorithm to improve the efficiency of the process and to include more distance functions. Although the literature lead to the assumption that the spatially distributed methods fit betters in satisfying user needs, there were not any hard evidence of that until the work of Tang and Sanderson [100]. In this work the authors proved the validity of such methods by performing a user preference study using Amazon Mechanical Turk[9]

In [71], Martins et al. proposed an approach based on the SVM learning to rank technique. The idea was to adapt the traditional machine-learning ranking approach to the combination of both thematic and geographic scores. This technique gives a good alternative to static combination methods, however its main drawback consists on the lack of strong training datasets.

Finally, there have been also some attempts to involve the user in the combination of multidimensional scores. These approaches [25, 46], delegate the final decision on the combination to the final user. They were based on visual representations of the document relevance in both the thematic and geographic analysis. Such approaches are effective from the point of view of the flexibility offered to the final user, but also because of this feature they present a limitation for non-expert users, or for processing large amounts of data.

### Discussion

The ranking combination techniques provided as part of this solution were designed merely based on experimentation. The implemented idea is based on combinations of thematics and geographic rankings (see 4.4). As stated previously the fact of having similar indexing structures for both the thematic and geographic aspects, makes easier the design of a valid combination strategy.

## 2.6 Evaluation of GIR systems

Has a new proposal actually improved already existing approaches? How is this improvement measured? GIR, like IR is an empirical discipline. The evaluation of a new approach is made by the design of experiments based on representative data collections. In general, the process includes two main tasks:

1. To choose an appropriate test collection that allows to verify if the new proposal can successfully solve the problems it is supposed to handle.

2. To select the metrics that measure how effective is the new proposal.

Both of these tasks are fundamental for obtaining an accurate estimation about what a new proposal can actually accomplish. Test collections should contain examples of problems that need to be solved and the metrics should bring reliable values that allow to compare the results among other approaches. Below a review of both, the evaluation metrics and the test collections that have been used by the GIR community last years.

[9] https://www.mturk.com/mturk/welcome

### 2.6.1 Test Collections

The goal of a test collection is to verify the ability of an approach to solve certain kind of problems. In GIR test collections follow the same structure that the ones built for traditional IR. In general, a test collection is composed by:

- A collection of documents.
- A set of user queries made according to this document collection.
- A set of relevance judgments, usually a binary assessment for each query-document pair.

Queries and documents should be carefully selected in order to ensure that the different types of scenarios targeted by the evaluation are included. Relevance judgments are usually made manually. For each query, a subset of the document collection is analyzed classifying each document in relevant or non-relevant. Evaluation metrics (see Section 2.6.2) in combination with this set of judgments allows complete the assessment process.

To the best of our knowledge, the first and only attempt to create large tests collection for evaluating entire GIR solutions has been the initiative of the GeoCLEF campaigns [40,41,68,69]. These campaigns designed challenge tracks oriented to the resolution of GIR problems. This initiative was organized by the Cross Language Evaluation Forum [10] between the years 2005 and 2008. This test collection was intended to comparatively evaluate systems on the basis of geographic relevance. The corpus is composed by a set of articles from *The Los Angeles Times (1994)* and *The Glasgow Herald (1995).* There are a total of 100 queries that were created along the four campaigns (25 queries each year). Furthermore this test collection is not publicly available, an author of a previous approach kindly provided the data for this work.

On the other hand, there are other test collections available that can be used for evaluating standard tasks within a GIR system. That is the case of collections dedicated to the evaluation of the toponym recognition and disambiguation processes. These ones do not follow the same structure presented above because they do not contain a set of queries, as their goal is to verify if the geographic references were properly identified. The following is a list containing some of the existing test collections for this purpose:

- TR-CoNLL: It is stated to be the first appropriate corpus for evaluating toponym recognition and disambiguation approaches. It was built by Leidner during his PhD research work [60]. The corpus is a subset of 946 documents from the CoNLL 2003 corpus, annotated with named entities. The resulting subset was then annotated with geocoordinates, resulting in a gold standard for geocoding containing 1903 annotated toponyms. Most toponyms in TR-CoNLL refer to countries, with the other toponyms referring to states and cities. The main limitation of this test collection is that it is not freely available.
- LGL: It is another test collection for evaluating toponym recognition and disambiguation algorithms. It was created by Lieberman, who named it Local-Global-Lexicon corpus [63]. This corpus is composed by 588 news

[10] CLEF[11]

items originating from all over the world. All news items are in English and were collected in 2009. They were manually annotated, including references to the Geonames knowledge base. Due to the characteristics of the corpus, LGL is more appropriate for evaluating geocoding systems on a local level. This test collection is not publicly available.

- GeoSemCor: It is presented by Buscaldi in [23]. It is built over the SemCor corpus, limited to its geographical names. It is also intended for evaluating toponym recognition and disambiguation. It is freely available and it contains 1210 toponyms but annotations are only toponyms existing in the WordNet ontology, making this test collection have a low coverage.

- TUD-Loc2013: It was created at the TU Dresden by Philipp Katz, David Urbansky, and Uliana Andriyeshyna [52]. Like the above mentioned, it was built for evaluating toponym recognition and disambiguation approachess. It is freely available containing 152 documents obtained from different pages from the web between 2012 and 2014. It has 3814 annotations, of which 90.51% include geographical coordinates.

Due to the characteristics of the test collections as well as their availability to be downloaded, the assessment process in this work is done by using the TUD-Loc2013 test collection in the evaluation of the topoynm recognition and disambiguation algorithms and GeoCLEF test collections (all four campaigns included) for evaluating the entire GIR proposal.

### 2.6.2 Evaluation Metrics

The evaluation metrics currently used in GIR are the same used for assessing traditional IR systems. These metrics are based on the the notion of relevant and non-relevant elements. The retrieval process is then evaluated according to the number of relevant and non-relevant elements returned in response to the user information need. In this sense, two well-known measures have been defined: precision and recall [79].

**Definition 2.6.1** (Precision)**.** Let *rel* be the set of relevant documents corresponding to a query and let *ret* be the set of relevant documents retrieved by the system. The *precision* function returns the fraction of retrieved documents that are relevant. More formally, it is defined as:

$$precision = \frac{|rel \cap ret|}{|ret|}$$

**Definition 2.6.2** (Recall)**.** Let *rel* be the set of relevant documents corresponding to a query and let *ret* be the set of relevant documents retrieved by the system. The *recall* function returns the fraction of relevant documents that are retrieved. The *recall* function is defined by the expression:

$$recall = \frac{|rel \cap ret|}{|rel|}$$

The unlikely perfect result is achieved when $recall = precision = 1$. This is the case where all and only the relevant documents are retrieved. Getting this perfect result is practically impossible, for example, usually in order to increase the recall, documents that are not actually relevant are also retrieved, which produces a decrease in the precision. The same happens in the other way around. Therefore, there is a trade-off between both functions.

A common metric that provides a balance between *precision* and *recall* values is the $F - measure$, which is the weighted harmonic mean of *precision* and *recall*.

**Definition 2.6.3** (F-measure)**.** Let $P$ and $R$ be the precision and recall values respectively. The $F - measure$ is defined as:

$$F = \frac{1}{\alpha \times \frac{1}{P} + (1-\alpha) \times \frac{1}{R}} = \frac{(\beta^2 + 1) \times R}{\beta^2 \times P + R}$$

where $\beta = \frac{1-\alpha}{\alpha}$ and $\alpha \in (0, 1]$

The default $F - measure$ is obtained when $\alpha = \frac{1}{2}$, $\beta = 1$, thus the *precision* and *recall* values are equally weighting. It is called $F_1$ measure and by substituting the $\alpha$ and $\beta$ values, the resulting formula is:

$$F_1 = \frac{2 \times P \times R}{P + R}$$

The above three measures are based on the list of retrieved documents. However, these do not consider the order of the elements in the list. For example, the evaluation of the toponym recognition and disambiguation processes only requires to verify if the retrieved elements are relevant or not. In this case, it is not important where the elements are ranked in the retrieved list, in fact, this list is treated as a set.

This is not the case for evaluating IR solutions where the relevance of the results counts. In general retrieval systems documents are ranked by their relevance according to an user information need. Users expect relevant documents in the first positions of the list. Therefore, in these cases the order matters and the evaluation should include the ability of the system of giving precedence to the relevant documents over less relevant ones. For this purpose, the Mean Average Precision ($MAP$) has become the standard metric and subsequently one of the most used for evaluating the retrieval and ranking process.

**Definition 2.6.4** (Average Precision)**.** Let $P(k)$ be the precision at cut-off $k$ in the ranked list obtained by the system in response to the query $q$. Let $relevant(k) \in \{0, 1\}$ be equal to 1 if the document at position $k$ in the ranked list is relevant and 0 otherwise. Let $rel$ be the set of relevant documents corresponding to the query $q$. The Average Precision is defined as:

$$AP(q) = \frac{1}{|rel|} \sum_{k=1}^{n} P(k) \times relevant(k)$$

**Definition 2.6.5** (Mean Average Precision)**.** Let $Q = \{q_1, q_2, \ldots, q_m\}$ be a set of queries. The Mean Average Precision ($MAP$) is the mean of the average precision values of each query. More formally, the $MAP$ function is defined as:

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{m} AP(q_i)$$

As it can be appreciated from the definitions, $AP$ constitutes the average precision value of each relevant document retrieved, divided by the total number of relevant documents in the collection, and $MAP$, given a set of queries, provides a single evaluation value of a system based on the average precision of each query in the set.

In [9], Baeza gives a more detailed description of these and other measures that allow to evaluate IR systems and consequently GIR systems. The evaluation metrics used in this work are $F_1$ and $MAP$. The former is used for assessing the toponym recognition and disambiguation algorithms while the latter is used for evaluating the entire GIR solution. The selection of these metrics has been also made aiming to establish a comparison with existing results reported in the literature for similar tasks.

### 2.6.3 Complete GIR systems

Although the work related to GIR components is quite extensive in the community, there is a lack of full end-to-end GIR solutions that allow the user to perform a complete GIR task.

There have been, though, some attempts to develop and provide such tools to the community. Perhaps the most prominent one, at the moment of writing this thesis, was the *SPIRIT* project citeJones04. In the literature it is possible to find references to other systems as well like *BUSTER* [106] and *GEOSEM* [14].

#### Discussion

One of the main goals of this work is to provide such a tool, with the inclusion of the base implementation for all the main components of a GIR system. In addition, the proposed GIR system, has been designed using a plugin mechanism that aims to ease the inclusion of new implementations for each one of the proposed components.

## 2.7 Conclusions

In this chapter we have reviewed some of the existing works in the areas that tribute to the design of a GIR system and its evaluation.

One of the key elements is the importance of relying on an external knowledge base [22,42,108]. As stated in the beginning of this chapter, this is common in scenarios targeting a particular domain. The depth of these knowledge bases depend on its complexity and the amount of information they provide. The kind of resource par excellence for this purpose are the semantic ontologies. Because of the domain size of geographical elements around the world, some of these semantic knowledge bases are also available in the form of traditional database [42, 108]. We have detected that among the existing resources, WordNet [76] and GeoNames [108] provides respectively an extensive knowledge on the English language and the Geographic places around the world. The custom knowledge base used in this work has been obtained by combining information from these two sources.

Another key conclusion that can be drawn from the related literature is the necessity of involving at least some kind of interpretation of the natural language. The problem is that although having an extensive knowledge base may allow to detect references to geographic places, the natural language poses a lot of ambiguities that cannot be resolved without a deeper understanding of the context. For this purpose, among the analyzed tools the Stanford CoreNLP toolkit from Manning [70] stands as one of the best alternatives. This tool, along with the information provided by an external knowledge base, play an important role in our toponym resolution strategy

We also reviewed the existing techniques for processing document and queries from a geographical point of view. The main goal in this process is to extract quantifiable information on the geographic references in documents and queries. In order to evaluate the similarity between them afterwards. On the analysis of the existing similarity measures, we have detected the importance of considering both the physical distance and the topological distance [8, 93]. Being the latter the distance that can be defined in a graph representing the domain of geographic elements according to some kind of hierarchy. Usually this hierarchy is defined according to the geo-political divisions of continents, countries, cities, etc.

We have also reviewed the existing resources available for evaluating GIR systems. We have found that there is in general a lack of large test collections for this purpose, the one built as part of the CLEF initiative [40,41,68,69] being the best option. On the other hand, there are other test collections that can be used for testing parts of a GIR solution, like the toponym resolution problem [52]. We remark that a common problem among the existing test collections is their availability. Usually the test collections for this purpose are not public which hinders the research attempts in this area.

Finally, we conclude this chapter by presenting the most commonly used evaluation metrics in the area of IR. In this work, we rely on the $F_1$ metric, for evaluating the toponym resolution process and on the $MAP$ metric for evaluating the performance of the overall solution.

# Chapter 3

# GeoNW: An ad-hoc geographic knowledge base

In Chapter 2 the usefulness of external resources in the development of GIR techniques has been mentioned. Figure 3.1 and Figure 3.2 shows the geographic process proposed in this work. As it can be seen, the process is divided in two main sub-processes: Geographic Information Extraction and Geographic Retrieval and Ranking. Most of the techniques included in both sub-processes are based on a geographic knowledge base which contains detailed information about the geographic references around the world. This resource, called GeoNW, constitutes a key aspect to obtain significant results in the geographic analysis process.

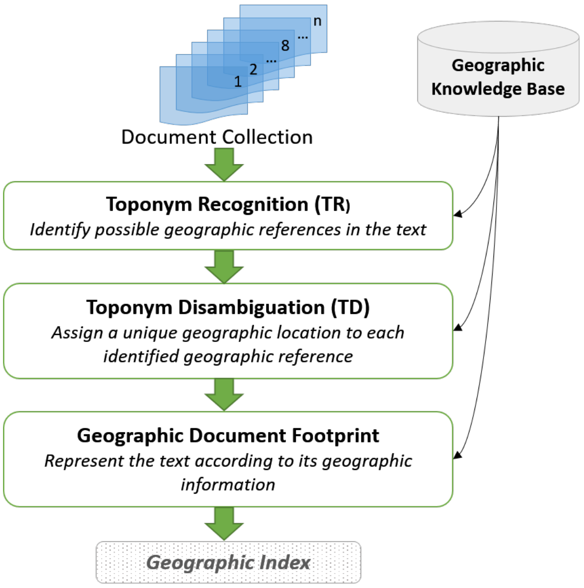
n
...
8
...
2
1
Document Collection
Geographic Knowledge Base
Toponym Recognition (TR)
Identify possible geographic references in the text
Toponym Disambiguation (TD)
Assign a unique geographic location to each identified geographic reference
Geographic Document Footprint
Represent the text according to its geographic information
Geographic Index

Figure 3.1: Geographic Information Extraction sub-process

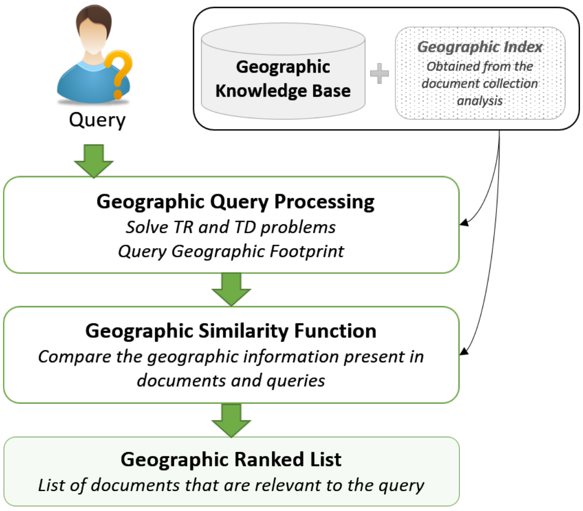
Query
Geographic Knowledge Base
Geographic Index
Obtained from the document collection analysis
Geographic Query Processing
Solve TR and TD problems
Query Geographic Footprint
Geographic Similarity Function
Compare the geographic information present in documents and queries
Geographic Ranked List
List of documents that are relevant to the query

Figure 3.2: Geographic Retrieval and Ranking sub-process

GeoNW has been built as the integration of two very well-known knowledge bases: WordNet and GeoNames. Although WordNet contains valuable lexical and semantic information that can be useful for solving some of the GIR problems, there are two main disadvantages in using it in geographic analysis scenarios: i) the small number of stored geographic references and ii) the lack of geographic coordinates related to each toponym [23]. On the other hand, GeoNames is very popular in the GIR community because it includes more than 10 millions of toponyms from all over the world with their corresponding geographic coordinates. However, this information is not as precise and accurate as required, producing a lot of ambiguity problems. Some attempts to solve these inconveniences have been proposed along the last years. As it was mentioned in Section 2.1, GeoWordNet [42] is one of those attempts. Like our approach, it has been built using WordNet and GeoNames, and also MultiWordNet[1] but the latter is mainly for adding support for the Italian language. The main drawback of GeoWordNet is that does not include the geographic information present in WordNet that is related to some specific geographic instance.

For example, Figure 3.3 and Figure 3.4 shows the geographic information related to the *Pacific Ocean* present in WordNet and GeoNames respectively. Since GeoWordNet does not consider WordNet instances, it only includes the information provided by GeoNames, dropping information existing in WordNet such as the description of the instance as well as its meronyms[2] can be useful for solving ambiguity problems. Even when there is short number of geographic instances in WordNet, these instances can be considered as the more used in web documents. For understanding better this idea, let's take a look into how WordNet is built [96, 97]. WordNet has been built by the development of some information extraction algorithms that run over large collections of data. This

[1] http://multiwordnet.fbk.eu

[2] A meronym is a semantic relation that denotes a constituent part of, or a member of something

data is composed by common documents from the Internet like Newswire corpora or Wikipedia. This allows to think that geographic instances in WordNet are those more likely to appear in documents from the Web.

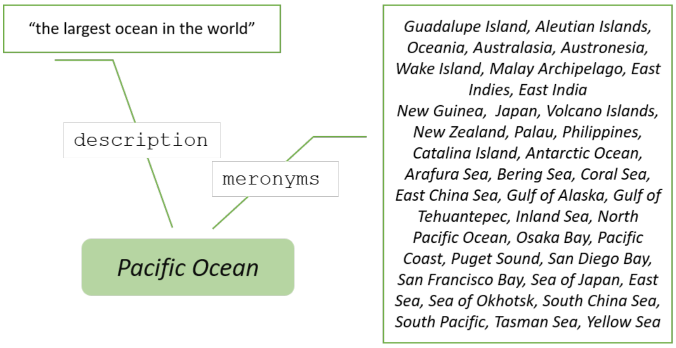
"the largest ocean in the world"
description
meronyms
Pacific Ocean
Guadalupe Island, Aleutian Islands, Oceania, Australasia, Austronesia, Wake Island, Malay Archipelago, East Indies, East India New Guinea, Japan, Volcano Islands, New Zealand, Palau, Philippines, Catalina Island, Antarctic Ocean, Arafura Sea, Bering Sea, Coral Sea, East China Sea, Gulf of Alaska, Gulf of Tehuantepec, Inland Sea, North Pacific Ocean, Osaka Bay, Pacific Coast, Puget Sound, San Diego Bay, San Francisco Bay, Sea of Japan, East Sea, Sea of Okhotsk, South China Sea, South Pacific, Tasman Sea, Yellow Sea

Figure 3.3: Geographic information related to "Pacific Ocean" within WordNet

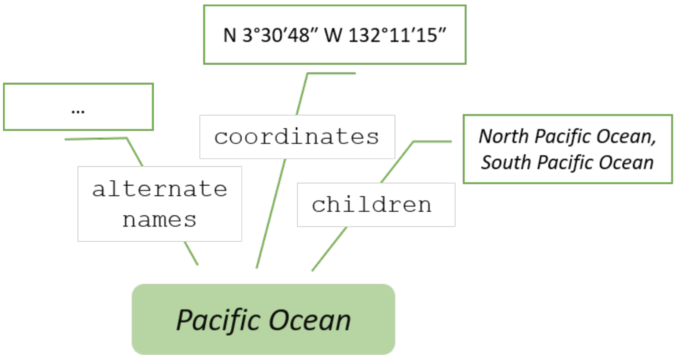

Figure 3.4: Geographic information related to "Pacific Ocean" within GeoNames

The rest of this chapter follows with an overview of GeoNames and WordNet describing its main characteristics as well as their advantages and disadvantages. Further sections present GeoNW as an ad-hoc geographic knowledge base, that stores and organizes the information obtained from both resources that could be relevant for solving well-known GIR problems.

## 3.1 Preliminaries

In spite of GeoNames issues, it is probably the most widely used resource in geographic text analysis. On the other hand, although WordNet is not a geographic knowledge base, it contains valuable semantic information that can contribute in the resolution of geographic ambiguities, which is a really challenging problem in GIR. Thus, which are the GeoNames main features that provide such relevant information for the geographic analysis? How can these GeoNames issues affect the results of a geographic analysis process? Which are the contributions of WordNet to solve GIR problems? Which are the disadvantages of using WordNet in the analysis of the geographic information in unstructured texts?

This section is intended to answer all of the above questions. It describes the structure and features (for geographic analysis purposes) of both, GeoNames and WordNet data sources. It explains their strengths and weaknesses for handling GIR problems while it leads to the assumption that an integration of both resources could produce a more robust geographic knowledge base.

### 3.1.1 GeoNames Structure and Features

The Figure 3.5 shows the scheme of the information contained in GeoNames. A *geoname* element represents a unique toponym: a place in the world, along with a series of attributes that conforms its geographic information. These attributes include: geographic coordinates, name, feature class/code, up to four levels of its political administrative ancestors, etc.

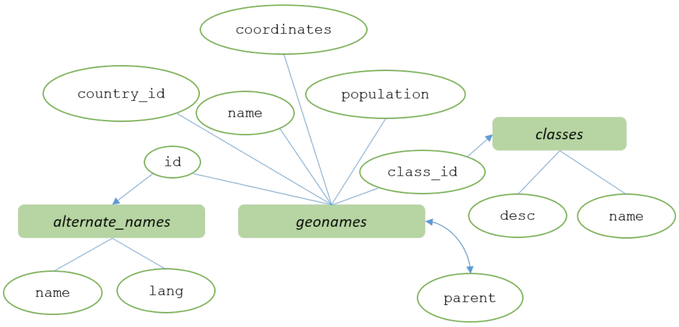

Figure 3.5: Geoname structure overview

The feature class/code information classify the element from a geopolitical or physical point of view (i.e. a capital city, a river, an historical place, etc.). The alternate names express all the known names for a place, this information may include names in different languages. Finally, GeoNames also provides a hierarchy that allows to express additional non-administrative relationships between the toponyms (i.e., *Pacific Ocean* contains *North Pacific Ocean* and *South Pacific Ocean* – see Figure 3.4).

Geonames also expresses relationships among its elements either explicitly or implicitly. The explicit relations available in GeoNames are:

- `is-part-of` this is a relationship among geoname instances, every element includes the information that links it to its direct parent in the political-administrative hierarchy.
- `is-synonym` this relation is defined between a place and its known names. It is represented by the link from alternate names to their corresponding geoname.
- `type-of` this is a relation between a geoname and its feature type. In GeoNames, this is a one to one relationship.

The implicit relations in GeoNames are:

- `sibling-of`: this information is recovered from the `is-part-of` relationship. It is defined by two geonames that share the same direct parent.

- `inside-of`: this information can be also built upon the `is-part-of` relationship across multiple levels. It is defined by a geoname and any of its ancestors (i.e., *Rome* is `inside-of` *Europe*).

- `contains`: this information is the inverse of the `is-part-of` relationship ($a$ `contains` $b$ if $b$ `is-part-of` $a$).

The following summarizes the main advantages and disadvantages of GeoNames for being used by GIR community:

- **Advantages**

  1. GeoNames coverage is one of the most wide-ranging among other existing geographic knowledge bases. It includes more than 10 millions of geographic instances distributed around the world.
  2. GeoNames provides the geographic coordinates of every geographic instance which allows to assign a unique geographic location to each toponym.
  3. GeoNames also includes explicit and implicit relationships among geographic instances (i.e., `is-part-of, contains, syblings, etc.`) that are very important for solving GIR tasks.

- **Disadvantages**

  1. The main disadvantage in GeoNames is the instance duplication problem. There are different instances that actually refer to the same geographic location. For example, there are 6 different instances exactly named *San Antonio* located in *California, US.* Table 3.1 shows the features corresponding to these 6 instances. It can be seen that 3 of these instances share exactly the same features. It means that are classified as `A.ADMD` which corresponds to administrative places and they are not only located in *California*, but also are contained in the same `admin2_code` hierarchy level. A further analysis over the geographic distance among these three places allows to think that all of them refer to the same place (Fig 3.6).

Table 3.1: Some of the geographic information in GeoNames corresponding to *San Antonio* places located in *California*

| GeoName Id | Feature Class | Feature Code | Country code | Admin1 code | Admin2 code |
|---|---|---|---|---|---|
| 5391623 | A | ADMD | US | CA | 001 |
| 5391624 | A | ADMD | US | CA | 001 |
| 5391625 | A | ADMD | US | CA | 001 |
| 5391626 | A | ADMD | US | CA | 037 |
| 5391627 | P | PPL | US | CA | 041 |
| 5391623 | A | ADMD | US | CA | 085 |

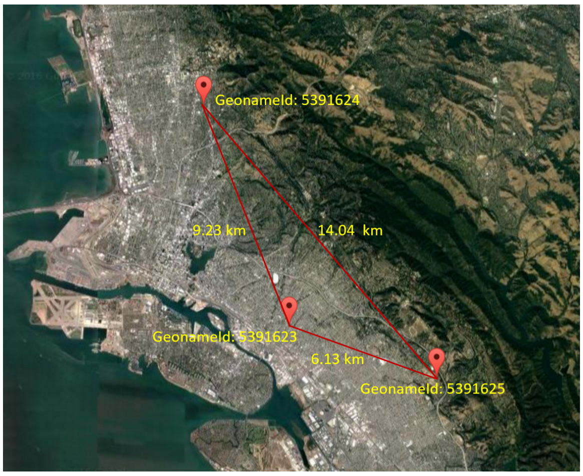

Figure 3.6: Geographic Distances among all 6 *San Antonio* places in *California* according to the geographic coordinates provided by GeoName

2. Another issue in GeoNames is the redundancy problem. Figure 3.7 shows part of the toponyms hierarchy corresponding to *Cuba*. The first redundancy is the presence of two places referring to *La Habana*. One of them with name *Havana* and class `P.PPLC` (capital city) and the other with name *La Habana* and class `A.ADM1` (first order administrative division). Both of these instances share their alternate names and their geographic coordinates are practically the same (about 10 km apart). In addition, the classes of these two elements can be considered equivalent as described in GeoNames documentation. Moreover, some of the children elements of these instances are also redundant, for example, the case of *La Habana Vieja*. The reason behind these similar classes and these redundancies are some failures in the integration of the different sources that conform GeoNames [109].

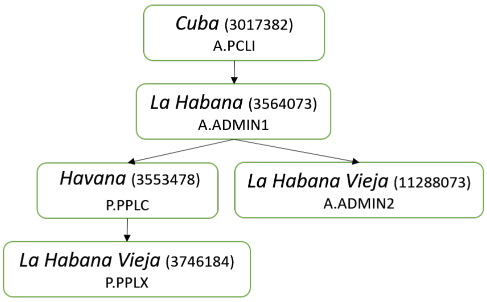

Figure 3.7: Part of the hierarchy in GeoNames corresponding to Cuba

3. GeoName classes also lead to other inconsistencies. For example, elements with low importance in the administrative classes are an-

cestors of elements of higher importance. A typical case of the above is Paris, see Figure 3.8 where *Paris*, the capital, classified as `P.PPLC` appears as a child of *Paris 04*, classified as `A.ADM4`.

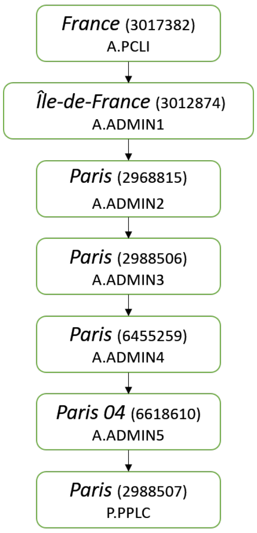

Figure 3.8: Part of the hierarchy in GeoNames corresponding to Paris in France

4. Some elements in GeoNames are wrongly located in the administrative hierarchy. For example, the city of *Sydney* in *Australia* `contains` *Bondi* neighborhood (Figure 3.9), however in GeoNames, *Bondi* appears as a child of *Bombala*, a `sibling` of *Sydney*. It is worth to notice that this kind of inconsistencies are very difficult to detect and fix in an automatic way.

5. Another disadvantage of GeoNames is the lack of a description related to the toponyms. Such a description can be very helpful in tasks like query expansion and toponym disambiguation.

Related to the first disadvantage previously mentioned one can think that ambiguity is also a problem in GeoNames. For example, if we extend the search for *San Antonio* to the whole world we would have found 2393 results. However this is not precisely a GeoName problem since, besides the duplication previously mentioned, there are thousands of places actually named *San Antonio*. As previously stated the presence of descriptions could considerable help in alleviating this.

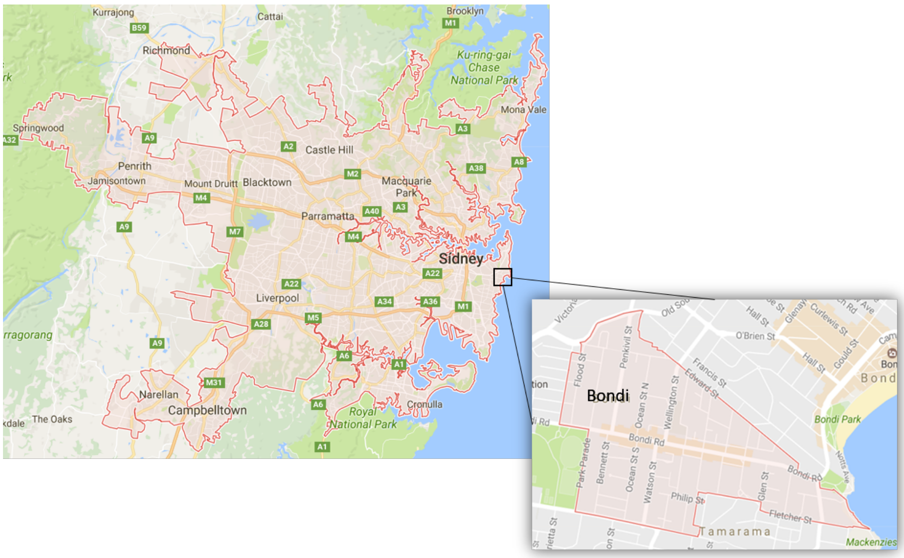

Figure 3.9: Geographic locations of Sydney and Bondi places. Notice the former `contains` the latter

### 3.1.2 WordNet Structure and Features

As it was previously mentioned, WordNet basic structure are the synsets. An example of synset is: (*Paris, City of Light, French capital, capital of France*). Each synset is connected to a unique `id` and a `description`. The description can be seen as the definition of the word. For example, the description of *Paris*'s synset is: the capital and largest city of France; and international center of culture and commerce). Synsets also are connected by different conceptual relationships (i.e. *Paris* synset is `holonym` of *Eiffel Tower, Louvre Museum, Montmartre*, etc.). The Figure 3.10 represents part of the scheme of the WordNet ontology[3].

Among the relationships expressed by WordNet the following are relevant for the geographic analysis:

- `synonymy`: it is considered the basic relation, because WordNet uses sets of synonyms (*synsets*) to represent word senses. It is a symmetric relation between instances sharing the same meaning. In the case of places all elements in the synset correspond to the same location.

- `meronymy`: this relationship expresses a *member-of* idea. For example, *Eiffel Tower* is a `meronym` of *Paris*.

- `holonymy`: This is the inverse of the `meronymy` relationship: *France* is an `holonym` of *Paris*.

- `denomymy`: it corresponds to the adjective used for inhabitants of a place (i.e. *Parisian* is the denomym of *Paris*)

[3] The presented scheme focuses on the geographic information contained in WordNet

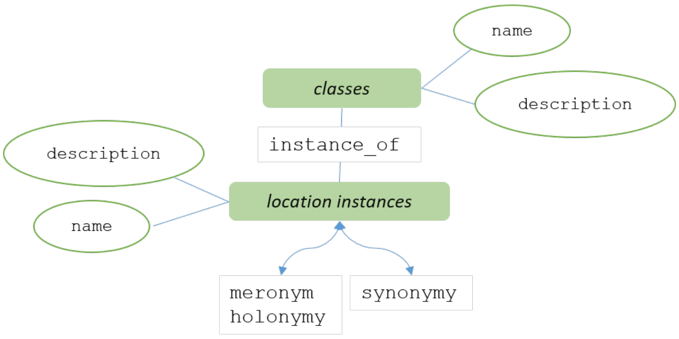

Figure 3.10: WordNet ontology structure overview with respect to geographic information

- `instance-of`: An instance is a representative element of a class. The *Black Sea* is an `instance-of` *Sea*.

The following summarizes the main advantages and disadvantages of WordNet with respect to its use as geographic knowledge base:

- **Advantages**
  1. WordNet provides rich descriptions for a certain number of toponyms, mainly countries and important cities. Because of how WordNet has been built, these place names can be considered those with higher presence in the Web.
  2. WordNet synsets include colloquial common references to certain places (i.e., *City of Light* is considered synonym of *Paris*). This enhances the detection of geographic references.
  3. WordNet provides information on location `demonyms` (i.e., *American*, *Parisian*, etc.) which also facilitates the detection of geographic references.
- **Disadvantages**
  1. WordNet coverage of locations around the world is fairly small when compared to other geographic knowledge sources like GeoNames. This limitation usually affects the recall in GIR tasks.
  2. WordNet does not contain information on the geographic coordinates corresponding to a toponym. This limits the possibility to reason about the distance between two places. Moreover, the geographic coordinates are very useful in several GIR tasks like toponym disambiguation, similarity functions and document scope detection.

## 3.2 GeoNW

Like most ontologies, GeoNW is composed by classes (also known as concepts), relationships among these classes as well as relationships among instances of these classes. GeoNW includes about 7 millions of instances that are the result of a filtering process on the GeoName ones (see Section 3.2.2). GeoNW contains 663 classes, which correspond mostly to the classes defined in GeoNames but enriched with the descriptions and relationships from the corresponding ones in WordNet (see Section 3.2.4). Like the classes, the instances in GeoNW are also enriched with a textual description. These descriptions can come either from WordNet or be automatically generated upon information already in GeoNW (see Section 3.2.3).

### 3.2.1 Structure and Features

Part of the classes and their hierarchical relations defined in GeoNW are shown in Figure 3.11. As it can be seen in the figure, rivers, lakes, mountains, deserts, etc. are considered physical places while continents, countries, cities and towns are considered administrative places.

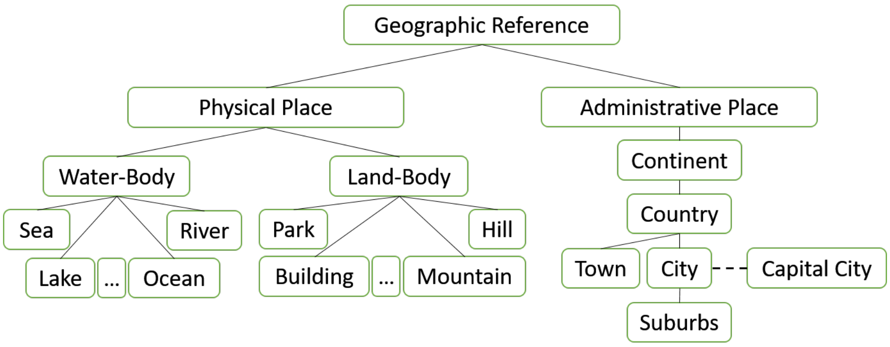

Figure 3.11: Some of the GeoNW classes and their hierarchical relations

The Table 3.2 summarizes the set of attributes and relationships defined on GeoNW instances. The first column in the table indicates the source of the information: *Generated* means the information has been somehow computed; *Hybrid* means the information is the result of merging GeoNames, WordNet and/or the automatic generated information.

| | **Attributes and Relationships** | **Description** |
|---|---|---|
| GeoNames | place_name | An attribute with the name of the place an instance is representing |
| | geo_coordinates | An attribute with the geographic coordinates corresponding to an instance |
| | population | An attribute with the population corresponding to an instance |
| | feature_type | An attribute that annotates an instance with its the corresponding class. It is composed by the feature_class and feature_code attributes in GeoNames. |
| | country_code | An attribute with the two letters code of the country where an instance is located. |

| | | |
|---|---|---|
| | `admin_level` | An attribute with the political-administrative information related to an instance. This relationship is obtained from the political-administrative hierarchy defined in GeoNames. |
| | `is-sibling` | A relationship that determines if two instances share the same immediate ancestor. This relationship is also obtained from the political-administrative hierarchy defined in GeoNames. |
| | `is-part-of` | A relationship that determines if an instance is part of another instance. A City `is-part-of` a Country, a Country `is-part-of` a Continent. From where, it can be inferred that a City `is-part-of` a Continent. This relationship is also obtained from the political-administrative hierarchy defined in GeoNames. |
| | `contains` | A relationship that determines if an instance contains another instance. It is also obtained according to the political-administrative hierarchy defined in GeoNames. A Country contains a City, a Continent contains a Country. From where, it can be inferred that a Continent contains a City. This relationship is also obtained from the political-administrative hierarchy defined in GeoNames. |
| **WordNet** | `non-admin-part-of` | A relationship that determines if one instance is part of another one without taking into account the political-administrative division. This relationship is obtained from the `meronymy` relationship in WordNet. It allows to relate instances that do not have an `is-part-of` relationship in GeoNames but they actually have a `part-of` relation according to their geographical location. For example, *Guadalupe Island* is `non-admin-part-of` *Pacific Ocean* |
| | `non-admin-contains` | A relationship that determines if one instance `contains` another one whithout taking into account the division political-administrative. This relationship is obtained from the `holonymy` relationship in WordNet. It allows to relate instances that do not have a `contains` relationship in GeoNames but they actually have a `contains` relation according to their geographical location.For example, *Pacific Ocean* `non-admin-contains` *Guadalupe Island* |

| | | |
|---|---|---|
| **Generated** | `is-distinctive` | The aim of this relationship is to define the most representative elements of an instance (location). For example, *Nice* is-distinctive of *France*, *Florence* `is-distinctive` of Italy, but *San Antonio* is not distinctive of *USA*, since there are a lot of *San Antonio* around the world and it is not a really representative place in *USA*. However, *USA* `is-distinctive` of the *San Antonio* located in *California* (see Figure 3.12). This relation is not symmetric, since for example, all countries are `distinctive` of their cities but not all cities are `distinctive` of their countries. By definition, the most populated children of an element are `distinctive` of it, for the particular case of countries the capital city is also considered `distinctive`. Such a relation allows to generate automatic descriptions for every location and then use these descriptions in the disambiguation process and also in the integration with other knowledge sources |
| **Hybrid** | `descriptions` | An attribute with a detailed description of an instance or class. This description can be obtained from GeoNames, WordNet or be automatically generated (see 3.2.3) |
| | `is-synonym` | A relationship that determines if one instance can also be referenced using another instance. This relationship can be imported from WordNet through the synsets and the `denonymy` relation or from GeoNames through the alternate names. |

Table 3.2: Attributes and Relationships in GeoNW

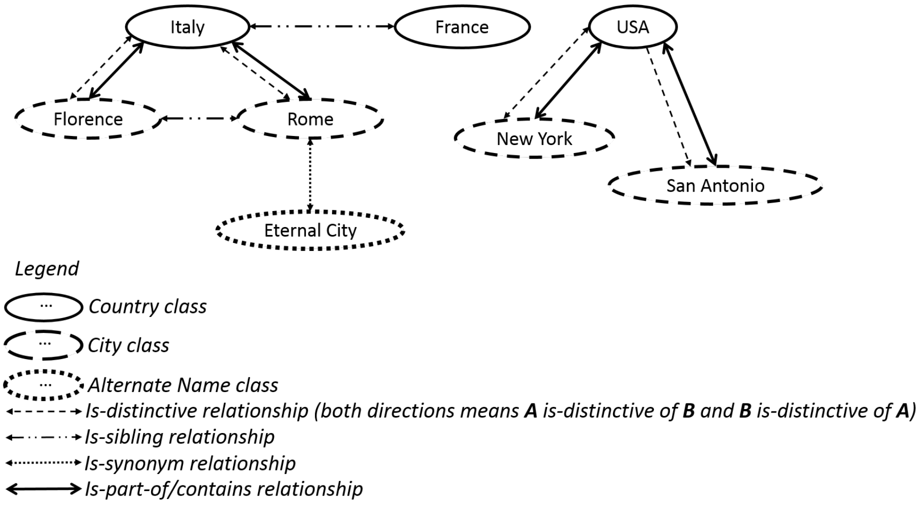

Figure 3.12: Some of the GeoNW instances and their relationships

### 3.2.2 Linking Process in GeoNames

Considering that Geonames is the most suitable public available geographic knowledge base, it is the main source of geographic information in GeoNW.

In Section 3.1.1 we refer to the duplication problem in GeoNames; and to the redundancy problem between elements of class `A` (country, state, region, etc.) and elements of class `P` (city, towns, villages, populated areas, etc.). Typically this kind of problem is solved by removing the elements causing the inconsistencies. However in order to keep the granularity of GeoNames instances we have decided to take a different approach. The idea is to add a `links-to` relationship between instances of these classes. The semantics of this relationship is to cluster elements that are duplicated or redundant. One of the elements in the cluster will be chosen as the representative element, the centroid. The rest of the elements will point to this centroid. This reduces future ambiguities since all elements in the cluster will be considered as one. For example, the duplicated elements from Table 3.1 will be linked in the following way:

Table 3.3: Linking of duplicated *San Antonio* places located in *California*

| **GeoName Id** | **Feature Class** | **Feature Code** | **Country code** | **Admin1 code** | **Admin2 code** | **Linked-to** |
|---|---|---|---|---|---|---|
| 5391623 | A | ADMD | US | CA | 001 | |
| 5391624 | A | ADMD | US | CA | 001 | 5391623 |
| 5391625 | A | ADMD | US | CA | 001 | 5391623 |

and for the redundancies example in Figure 3.8 and Figure 3.7 the linking process delivers the following result:

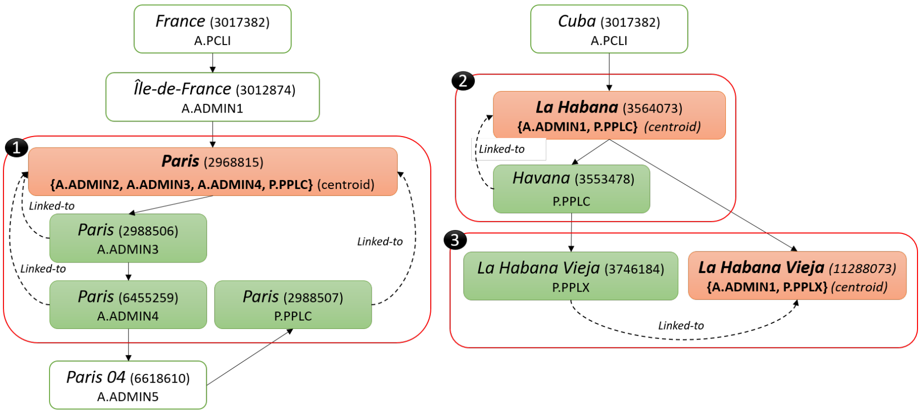

Figure 3.13: Final result of the linking process

In Figure 3.13 clusters are represented by bounding boxes, the elements in orange represent the centroids. We notice how the representative element summarizes part of the information from the other elements in its cluster.

### Linking algorithm

The linking algorithm follows a clustering strategy. Clusters' centroids will be the higher positioned element in the political-administrative hierarchy. In the case this is non-deterministic the one with the highest population among them is chosen. The centroid summarizes part of the information from its linked elements, specifically, the feature classes and the alternate names.

Below we present a set of definitions necessary for the description of the linking process, followed by the main steps of the algorithm. In these definitions we use the terms $t, t_i$ to refer to toponyms; we access to attributes of these toponyms by writing $t.attribute$ (i.e., $t.name$, $t.parent$); cluster elements are represented as $C, C_i$.

**Definition 3.2.1** (Names)**.** Let $t$ be a toponym in GeoNames, the function *names* gives the set of all known names of $t$: its current name plus its alternate names. More formally:

$$names(t) = \{t.name\} \cup \{x : x \in t.alternate\_names\}$$

**Definition 3.2.2** (Path)**.** Let $t_0$ be a toponym in GeoNames, the function *path* returns the sequence of all ancestors of $t_0$. More formally:

$$path(t_0) = \{t_1, t_2, \ldots, t_n\}$$

where $t_{i+1}$ is the immediate parent of $t_i$ and $t_n$ corresponds to the root element in the hierarchy.

**Definition 3.2.3** (Parent*)**.** Let $t$ be a toponym in GeoNames, the function *parent** returns the corresponding parent of $t$:

$$parent^*(t) = \begin{cases} t.parent.centroid & \text{if } t \text{ is a centroid} \\ t.centroid.parent.centroid & \text{otherwise} \end{cases}$$

**Definition 3.2.4** (Equally Named Ancestors)**.** Let $t_1$, $t_2$ be two toponyms in GeoNames. $t_1$ and $t_2$ are *linkable* iif:
$(i)$ $t_1.class = t_2.class$ or $t_1.class, t_2.class \in \{\mathtt{A}, \mathtt{P}\}$ and
$(ii)$ $t_1.name = t_2.name$ and
$(iii)$ $t_1.country\_code = t_2.country\_code$ and
$(iv)$ $t_1 \in path(t_2)$ or $t_2 \in path(t_1)$
where $A$ and $P$ are classes defined in GeoNames that classify a toponym as administrative or populated place respectively.

**Definition 3.2.5** (Coincident Alternate Names)**.** Let $C_1$ and $C_2$ be two clusters of one or more elements, satisfying the Definition 3.2.4. Let $t_1$ and $t_2$ be the centroids of $C_1$ and $C_2$ respectively. $C_1$ and $C_2$ are *linkable* iif:
$(i)$ $t_1.class = t_2.class$ or $t_1.class, t_2.class \in \{\mathtt{A}, \mathtt{P}\}$ and
$(ii)$ $\{\bigcap names(x_i) : x_i \in C_1\} \cap \{\bigcap names(x_j) : x_j \in C_2\} \neq \emptyset$ and
$(iii)$ $t_1.country\_code = t_2.country\_code$ and
$(iv)$ $t_1 \in path(t_2)$ or $t_2 \in path(t_1)$
where $A$ and $P$ are classes defined in GeoNames that classify a toponym as administrative or populated place respectively.

**Definition 3.2.6** (Equally Named Siblings)**.** Let $C_1$ and $C_2$ be two clusters of one or more elements, satisfying the Definition 3.2.4 or Defintion 3.2.5. Let $t_1$ and $t_2$ be the centroids of $C_1$ and $C_2$ respectively. $C_1$ and $C_2$ are *linkable* iif:
$(i)$ $t_1.class = t_2.class$ or $t_1.class, t_2.class \in \{\mathtt{A}, \mathtt{P}\}$ and
$(ii)$ $t_1.name = t_2.name$ and
$(iii)$ $parent^*(t_1) = parent^*(t_2)$
where $A$ and $P$ are classes defined in GeoNames that classify a toponym as administrative or populated place respectively.

The algorithm is divided in three steps:

**Step 1:** All the elements that satisfy the Definition 3.2.4 will be clustered. At the end of this first process, elements with exactly the same name, that are located in the same country and are contained in the same path to the root in the political-administrative hierarchy, will be included in the same cluster. Each element in the cluster will be `linked-to` the centroid which is selected as the one at the highest position in the political administrative hierarchy. In this way, it is solved one of the above mentioned problems in GeoNames. Figure 3.13, the cluster number 1, shows how related *Paris* instances are linked in GeoNW avoiding the ambiguity that can affect the results of further GIR processes.

This `linked-to` relationship allows to consider a new parent (see Definition 3.2.3) for each non-centroid element in the cluster without altering GeoNames original structure. The idea behind this approach is to preserve the granularity level provided by GeoNames but reducing toponym ambiguity.

**Step 2:** A second step will continue the clustering process attending to the criteria of the Definition 3.2.5. This step will take into account those elements that share at least one exactly equal alternate name. Figure 3.13, with the inclusion of the cluster number 2 another of the GeoNames issues is solved. For our purposes, *La Habana* and *Havana* are referring to the same place. The definition can be applied for single elements or for elements that are already centroids in their corresponding clusters.

**Step 3:** The clustering process conclude by relating all the elements that satisfy the Definition 3.2.6. Previously steps define the case where two elements in the same path to the root can be considered the same, according to our purposes. This last step allows to link elements that share their name and are siblings. The addition of the cluster number 3 ( Figure 3.13), allows to consider that both *La Habana Vieja* instances are referring to the same place. Like in the **Step 2**, the definition can be applied for single elements or for elements that are centroids in their corresponding clusters.

The clustering process deals with the duplication and redundancy problems previously mentioned. Notice that during the process there is not any modification over the information obtained from GeoNames. The process just adds new information that allows to relate geographic instances in new ways (Figure 3.13). In summary, GeoNW keeps the information present in GeoNames but their elements are clustered and represented by its corresponding centroid which includes all the features of its cluster. Notice that this process brings three major changes in the toponyms structure. The addition of a `linked-to` relationship, the addition of a `parent*` relationship and the possibility for a toponym to belong to more than one feature class.

Unless specified otherwise in the rest of this document when we refer to GeoNW toponyms we are referring to those that remained centroids after the linking process. The whole goal of this modification has been to alleviate toponym ambiguities from GeoNames. After this linking process about 2 millions of elements have been linked to a representative element. We can infer from the description of this process that the linked elements were in every case elements causing either duplication or redundancies.

### 3.2.3 Automatic generation of toponyms description

The automatic generation of toponyms description aims to enrich the information related to geographic instances. It will further contribute to solve toponym ambiguity problems in scenarios where the geographic analysis of unstructured texts is required. The Figure 3.14 shows the description of *La Habana*. This description is automatically generated by combining in a tuple information obtained from the attributes and relationships of an instance.

**Definition 3.2.7** (Names*)**.** Let $t$ be a toponym in GeoNW, and $C$ be a cluster such that $t$ is its centroid. The function *names** gives the set of all possible names of $t$. More formally:

$$names^*(t) = \{x : x \in \bigcup names(x_i) \ \forall x_i \in C\}$$

**Definition 3.2.8** (Classes Description)**.** Let $t$ be a toponym in GeoNW, and $C$ be a cluster such as $t$ is its centroid. The function *classes** gives the set of all the classes from which $t$ can be classified with their corresponding descriptions. More formally:

$$classes^*(t) = \{class : class \in \bigcup x_i.class \ \forall x_i \in C\}$$

**Definition 3.2.9** (Is-Distinctive Relationship)**.** Let $t_1$ and $t_2$ be two toponyms in GeoNW. $t_1$ `is-distinctive` of $t_2$ iif:

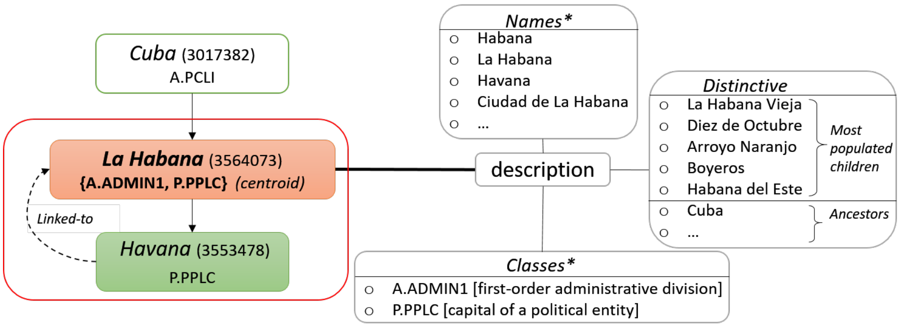

Figure 3.14: Final result of the linking process

$(i)$ $t_1 \in path(t_2)$ or<br>
$(ii)$ $parent^*(t_1) = t_2$ and
$(\texttt{P.PPLC} \in classes^*(t_1)$ or $t_1.population \geq 20\% * t_2.population)$

Where `P.PPLC` is the class in GeoNW for capital places.

**Definition 3.2.10** (Distinctive Function)**.** Let $t$ be a toponym in GeoNW. The function *distinctive* returns the set of elements that are distinctive for $t$. More formally:

$$distintictive(t) = \{t_i : t_i \texttt{ is-distinctive } of\ t\}$$

Notice that $t$ can be a single element where it is considered the only element of a cluster and subsequently its centroid; or $t$ can be the centroid of a cluster with more than one element.

**Definition 3.2.11** (Description)**.** Let $t$ be a toponym (and a cluster centroid) in GeoNW. The automatic *description* of $t$ will be a tuple defined as:

$$description(t) = \langle names^*(t), classes^*(t), distinctive(t)\rangle$$

The process of assigning automatic descriptions to toponyms takes place after the linking process. We remark that these descriptions are assigned to centroid elements. Thus they will summarize the information about other elements in its cluster. Intuitively, the idea is to be able to proceed considering only centroid elements, but making them contain any relevant information from their cluster.

### 3.2.4 WordNet Integration into GeoNW

A major challenge in our work was the integration of the information from WordNet into GeoNW. This integration faces two problems, the integration of the geographic instances and the integration of the concepts/classes.

#### Geographic Instances Integration:

At this point the information in GeoNW corresponds to the information in GeoNames except for the linking process previously described, and the addition of the automatic descriptions. The main goal of the integration process is to enrich GeoNW with information in WordNet related to geographic instances. To the best of our knowledge, this is the first attempt to integrate geographic instances from GeoNames and WordNet. Previous works integrate GeoNames classes with WordNet classes [42] or add geographic coordinates to the geographic elements in WordNet [21] but no one integrates the information of particular instances.

The basic idea of the integration is to look for equivalent geographic instances in WordNet and GeoNW. Each geographic instance in WordNet has specific descriptions that characterized a place. On the other hand, each toponym/centroid in GeoNW has a generated description (see Section 3.2.3). Therefore, the aim is to merge the elements with the most similar descriptions.

The similarity between the description of an element in WordNet and the generated description of an element in GeoNW is made using the well-known cosine similarity metric [90]. A first step ranges for each location in WordNet searching for all the elements in GeoNW with a coincident name. The best GeoNW candidate is chosen according to the similarity of the descriptions. The information from WordNet is merged into the GeoNW element description. More specifically:

- the elements in the WordNet location synset and its demonyms are added to the alternate names attribute

- the location synset definition in WordNet is added to a new attribute, that will also become part of the automatic description of GeoNW elements

This process is propagated to the children of the WordNet location (the elements in the `part-meronyms` WordNet relationship). Each child is merged with the best GeoNW candidate. There are two possibilities, the GeoNW element merged with the original WordNet location is an ancestor of the GeoNW element corresponding to the child, or not. In the second case, we update the GeoNames enforcing the same hierarchical relation existing in WordNet for that pair of elements. We notice that this is a very rare case and happens only for elements that do not correspond to political-administrative hierarchies, Figure 3.3 and Figure 3.4 are an example of this situation. In this case the information about the elements contained by *Pacific Ocean* is richer in WordNet than in GeoNW.

To better understand the process, let us think in the instance *Havana*. There is one single entry corresponding to this place name in WordNet. Instead, GeoNW contains more than 15 places with this name. The following corresponds to some of the descriptions of *Havana* in both WordNet and GeoNW:

***Havana* from WordNet** : capital of Cuba; Cuban capital; the capital and largest city of Cuba; located in western Cuba; one of the oldest cities in the Americas.

***Havana, Cuba* from GeoNW** : Habana, La Habana, Havana, Ciudad de La Habana. La Habana Vieja, Diez de Octubre, Arroyo Naranjo, Boyeros, Habana del Este, Cuba, North America; first-order administrative division; capital of a political entity.

***Havana, Illinois* from GeoNW** : Havana, Mason County, Illinois, United States, North America; seat of a second-order administrative division.

***Havana, Florida* from GeoNW** : Havana, Hillsborough County, Florida, United States, North America; populated place.

***Havana, Texas* from GeoNW** : Havana, Hidalgo County, Texas, United States, North America; populated place.

The Table 3.4 shows the results of the similarity values corresponding to the *Havana* instances in GeoNW. As it can be seen the cosine similarity function correctly determines the most similar descriptions with a value of 0.32. The Figure 3.15 shows (in red) the new information added to the toponym *La Habana* in GeoNW after merging it with the corresponding WordNet entry.

Table 3.4: Results obtained from the comparison between the geographic instance *Havana* in WordNet and its corresponding candidates in GeoNW, using the cosine similarity function

| **Instance comparison** | **Cosine Similarity** |
|---|---|
| sim(*Havana* from WordNet, *Havana, Cuba* from GeoNW) | **0.322** |
| sim(*Havana* from WordNet, *Havana, Florida* from GeoNW) | 0.200 |
| sim(*Havana* from WordNet, *Havana, Texas* from GeoNW) | 0.200 |
| sim(*Havana* from WordNet, *Havana, Illinois* from GeoNW) | 0.173 |

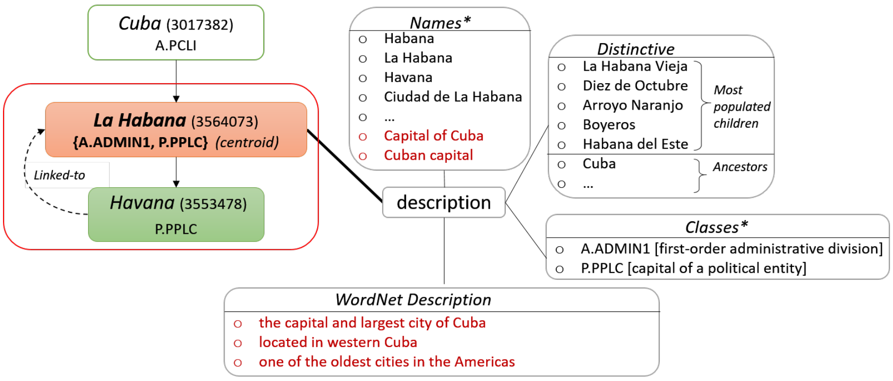

Figure 3.15: Result of including *Havana* WordNet information in GeoNW

**Definition 3.2.12** (Integration Condition)**.** Let *geo_wn* be a geographic instance (toponym) in WordNet. Let $t$ be a toponym in GeoNW. *geo_wn* is merged with $t$ iif:

$(i)$ $synset(geo\_wn) \cap names^*(t) \neq \emptyset$ and

$(ii)$ $cos\_sim(geo\_wn, t) = \max_i\{cos\_sim(geo\_wn, t_i)\ \forall t_i \in GeoNW\}$

where $synset(geo\_wn)$ is the set of all synonyms of *geo_wn* in WordNet and *cos_sim* is the cosine similarity function.

The above Definition 3.2.12 formally describes the instance integration process. As part of this process 6833 instances in GeoNW were enriched with information obtained from WordNet. We noticed that among the enriched elements there were all countries and their capital cities.

### Classes Integration

This process is based on the instance integration process. Intuitively, if two elements are considered the same, there is a high probability of their corresponding classes being also related. From the GeoNW structure, we have that a centroid toponym has associated the classes of all the elements in its cluster. When a match between a WordNet instance and a GeoNW instance is detected, each class in GeoNW related to the actual toponym is considered a candidate match for the class in WordNet that corresponds to the geographic instance that is being analyzed.

At the end of the instances integration process, we obtain a set of possible matches between GeoNW classes and WordNet classes. We select the best match among these by following the same idea as with the instances. The class description in GeoNW is compared with the class description in WordNet using the cosine similarity function. The couples with the highest results are then merged into GeoNW. In this case the merge takes the class description already existing in GeoNW and adds to it the description of the corresponding class in WordNet.

## 3.3 Conclusions

In this chapter we analyze the design of a custom new geographic knowledge base, called GeoNW. As described in Chapter 2 there are a number of existing resources for this purpose. Precisely, GeoNW has been built by integrating two of these already existing resources: WordNet and GeoNames.

We analyzed the main advantages and disadvantages of these two resources concluding that a combination of the main strengths of each one could be combined into a new resource. Actually, this idea has been considered before [42]. However, in the cited work, geographic instances in WordNet are dismissed, while we consider that those enclose important information not present in GeoNames. We remark that even though WordNet contains only a very limited number of geographic instances, the construction of WordNet ensures that commonly used elements in web documents like countries, capital cities, and important landmarks are included.

The design of GeoNW (partially) takes care also of some of the main problems in GeoNames like the duplication and redundancy of the information. This goal is achieved through the application of a linking process were elements considered equivalent are grouped into a single cluster and represented by its centroid. After this linking process the number of effective elements in GeoNW is reduced in approximately two millions of elements compared to the original number in GeoNames.

One of the main contributions of GeoNW with respect to GeoNames is the addition of a description for every place. This description is at first instance generated automatically based upon implicit or explicit information present in GeoNames. Such description has two main purposes. The first is to aid in the integration of the WordNet data, and the second one will be exploited by the toponym resolution process. In a second step the description of the elements also present in WordNet is enriched by adding the corresponding synset description.

The most complicated step in the construction of GeoNW is without doubts the integration of the information coming from WordNet. As stated at the beginning of this chapter a key information missing from WordNet are the geographical coordinates associated to location instances. This complicates the process of detecting ambiguities in GeoNW, where one name corresponds to the instance present in WordNet. This problem is solved by choosing the possibility with highest similarity between the synset description in WordNet and the description generated in GeoNW. The similarity measure used for this purpose was the well-known cosine similarity metric. Besides the already mentioned descriptions, the integration of WordNet brings into GeoNW additional information on alternative names and hierarchical relationships. It also allows to enrich the description of GeoNW instances.

# Chapter 4

# A set of techniques for GIR based on GeoNW

The previous chapter describes GeoNW, a new geographic knowledge base that aims to support the performance of general GIR systems. GeoNW plays a very important role in the development of different techniques involved in the accomplishment of a successful GIR process. Figure 3.1 and Figure 3.2 show the complete geographic analysis process. As it can be appreciated, GeoNW is used for solving most of the problems related to it. This chapter explains each of the techniques based on GeoNW that are proposed in this work. In summary, the chapter includes:

- Toponym Recognition (TR) algorithm: Its input is an unstructured text (query/document) and it returns the list of elements contained in the text that are considered geographic references.

- Toponym Disambiguation (TD) algorithm: Its input is the list of possible toponyms that was obtained from the TR algorithm.It returns a subset of this list, including for each toponym its geographic coordinates.

- Geographic Document Footprint: Having the geographic information related to a document. This process is intended to represent it and to quantify its relevance by proposing a geographic weighting model technique.

- Geographic Query processing: Its input is an unstructured text (user query) and it returns a representation of its geographic information equivalent to the one defined for the documents.

- Geographic Similarity Measure: Its input is a document and a query geographic representations and it returns their similarity value.

- Combination Ranking Strategy: Given two ranked list of documents (textual and geographic), it combines both lists properly returning a final ranked list of documents which is the final response of the user query.

## 4.1 Toponym Resolution in Unstructured Texts

The first step in every geographical analysis of unstructured texts is to identify the geographic locations. As it was mentioned in Section 2.5, because of the existence of several non-geo/geo and geo/geo ambiguities, to properly identify the geographic locations is a challenging task.

The TR process constitutes the very first estimation of the geographic references present in the text. Its goal is to determine if a word or sequence of words can be considered a toponym or not. That is why, it is intended to solve the non-geo/geo ambiguity. Ideally, the TR algorithm should provide the list of all toponyms in the text and the TD algorithm should assign to each toponym its corresponding geographic location. In practice, TR aims to reach a high recall leaving the precision to the TD algorithm. Both, TR and TD processes are fundamental for properly facing the further geographic retrieval steps.

### 4.1.1 Toponym Recognition (TR)

Like most of the toponym recognition techniques, our proposal is based on external resources, specifically it is based on GeoNW and Stanford CoreNLP tools. GeoNW contains about 8 millions of geographic locations around the world. This vast coverage has a positive and a negative effect for developing any TR strategy. The positive side: it is possible to recognize all the geographic references in the text. The negative side: there are several non-geo/geo ambiguities. On the other hand, CoreNLP includes the Named Entity Recognition (NER) tool that allows to recognize with a very high precision the geographic references in the text, however this tool is trained using rather small gazetteers losing a lot of place names that actually correspond to geographic locations. This analysis allows to think that a good approach can be to get a balance of both resources. In this way emerges our proposal.

#### CoreNLP tool

Stanford CoreNLP [70] is a Java framework that allows to process unstructured text. It contains most of the common core Natural Language Processing (NLP) techniques for annotating text (see Figure 4.1). Due to its simple utilization and its success in the annotation task, Stanford CoreNLP has become a widely used toolkit.

The Figure 4.1 shows the execution pipeline corresponding to the annotation process. The pipeline is composed by a set of tools that classify each word (i.e., part-of-speech tagger) as well as determine the relationship among these words (i.e., co-reference resolution). The present work uses two of these tools:

- **part-of-speech (POS) tagger**: It labels each word in the text with its grammatical category[1] [`noun, adjective, adverb, etc.`]. In [102], Toutanova explains in detail the whole pos-tagger process showing an accuracy of 97.24%.
- **named entity recognition (NER)**: It recognizes sequences of words in a text which are names of things. These sequences of words can be classified

[1] A complete list of all categories can be found here: `http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html`

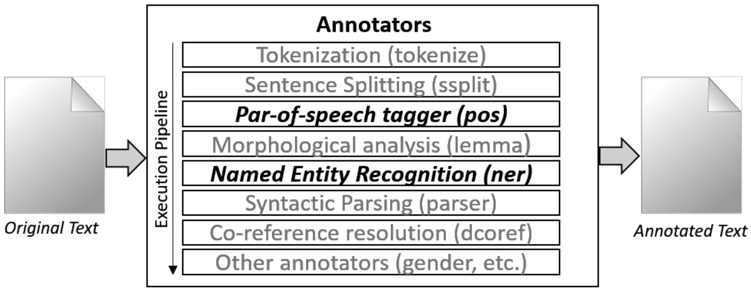
Annotators
Tokenization (tokenize)
Sentence Splitting (ssplit)
Par-of-speech tagger (pos)
Morphological analysis (lemma)
Named Entity Recognition (ner)
Syntactic Parsing (parser)
Co-reference resolution (dcoref)
Other annotators (gender, etc.)
Execution Pipeline
Original Text
Annotated Text

Figure 4.1: Stanford CoreNLP architecture. In bold the techniques used in this work.

in one of the following categories: [`location, organization, person, date`] (this actually depends on the training model). The remaining words in the text are skipped. In other words, they are not considered entities. In [35], Finkel presents the strategy followed for the recognition process. The proposed technique improves previous information extraction systems in up to 9%.

### The strategy

A document is processed by three tagging criteria: i) GeoNW; ii) POS tagger; and iii) NER tagger. The `null` classification is allowed and it means that the tagger did not recognize the element in any of its categories.

The result of each tagger is then combined for deciding whether the analyzed element is a geographic reference or not. In practice, we consider that a geographic term is a location when we are in presence of one of the combinations from Table 4.1.

The accuracy level shown in the table means that the toponym that satisfies the combination rule has a `High/Low` probability of being an actual geographic reference. This feature is used in the further TD analysis.

The first combination is the most obvious, elements found within `GeoNW`, and also classified as a location by `NLP.ner` tool have a `high` probability of being locations.

The second and third combination are a direct result from the experiments. An element recognized by GeoNW and classified as `Organization/ Person` by `NLP.ner` is has a `high` probability of being a location if the element name contains its corresponding `feature_class` name or the element is preceded/followed by another noun also recognized as location by GeoNW.

The fourth combination includes the case where an element is considered a location by GeoNW but its name does not contain its `feature_class` name, the `NLP.ner` tagger classifies it as `Organization/Person` and analyzing the `NLP.pos` tagger the element is not preceded or followed by a `noun`. Although this combination provides toponyms with a low probability of being actual locations, based on the experiments, it often produces correct detections. This fact, depends highly on the training data set of the NLP tools.

Table 4.1: Combinations of the three taggers that lead to an element being considered as a location. (`Anything` means that the tag is ignored by the defined rule and `Null` means that the tagger skipped the element.)

| **Rule #** | `GeoNW` | `NLP.ner` | `NLP.pos` | **Accuracy Level** |
|---|---|---|---|---|
| I | `Location` | `Location` | `Anything` | `High` |
| II | `Location` + `feature_class` name contained in the toponym name | `Organization/ Person` | `Anything` | `High` |
| III | `Location` | `Organization/ Person` | is a `noun group` preceded/ followed by a `noun` that is also recognized as a location in GeoNW | `High` |
| IV | `Location` | `Organization/ Person` | is a `noun group` NOT preceded/ followed by a `noun` | `Low` |
| V | `Location` + `feature_class` name contained in the toponym name | `Null` | `Anything` | `High` |
| VI | `Location` | `Null` | is a `noun group` preceded/ followed by a `noun` that coincides with the toponym `feature_class` name in GeoNW | `Low` |
| VII | `Location` | `Null` | is a `noun group` NOT preceded/ followed by a `noun` | `Low` |

The fifth combination also detects an element with a high probability of being a location. It takes into account an element that has been detected by GeoNW and its name contains the corresponding `feature_class` name. The result of this simple rule fully depends of the performance of GeoNW. Experiments shows very good results correctly recognizing as locations almost all the analyzed texts.

The sixth combination is also a direct result from the experiments. In this case, there is a low level of accuracy, but in many cases elements detected by GeoNW that are preceded/followed by its corresponding `feature_class` name are actually locations.

Finally, the seventh combination includes those elements recognized as locations by GeoNW but are not classified as an entity by `NLP.ner` and also it is not preceded of followed by any other noun. This is not a very reliable rule because it roughly produces a half of wrong classifications but we consider that a more exhaustive analysis is done by the TD strategy.

Table 4.2: Examples of failed and successful classifications obtained from the defined set of rules.

| **Rule #** | **Text** | **Classification** |
|---|---|---|
| I | There are a lot of amazing things to do in *Paris*. | |
| I | Most people believes that *Paris* killed Achilles. | Wrong |
| I | The recommendation came from *Austin* | Right or Wrong? |
| II | We arrived at *John F. Kennedy Airport* | Right |
| II | A famous collection from the *Library of Congress* in *Washington* | Right |
| II | An interview of the actor playing *Richard Castle* character. | Wrong |
| III | The concert was in *Congress, Ohio* | Right |
| IV | The concert was in *Congress*. | Right |
| IV | The *Congress* decided to vote against the proposal. | Wrong |
| V | The *capital of Cuba* is located in the *Caribbean*. | Right |
| VI | We already booked in the *Four Seasons Buenos Aires hotel*. | Right |
| VII | The novel of Jack *London* | Right |
| VII | The film was named *Red* | Wrong |

Table 4.2 shows some examples of failed and successful classifications resulted after applying the set of rules defined above. The column named **Classification** indicates whether the recognition result obtained from proposed TR algorithm fails or not. Notice how cases like *Austin* can be ambiguous even for human analysis. All the examples in the table are recognized as toponyms except the case of *Jack London* where the proposed TR algorithm successfully determines that *London* is not a location in that context. It is worth noting that these are preliminary results. All of these toponyms will be post processed through the TD algorithm. Combinations that provide a low level of accuracy have

been considered in order to increase the recall. In this way, the TD strategy is intended to determine with higher precision whether or not these toponyms are actually locations and to assign them the corresponding geographic coordinates.

### 4.1.2 Toponym Disambiguation (TD)

From the recognition process, we have a list of all possible toponyms identified in the text. Among other characteristics, each element of this list can be in one of the following cases:

1. The element is an unambiguous geographic location (e.g., *United States of America*). This is the best case and it does not require any post-analysis. We can directly assign to the element its corresponding geographic coordinates.

2. The element is a geographic location with a geo/geo ambiguity problem (e.g., *London, Ontario* or *London, UK*). In this case, the element refers to different locations around the world. It requires a deeper context analysis in order to determine which is the actual location it is referring.

3. The element is a geographic location with a geo/geo feature type ambiguity problem (e.g., *Mississippi* the river or the US state). In this case the element can refer to a physical geography (e.g., river, lake, mountain, etc.) or to an administrative geography (e.g., country, city, town, etc.). It also requires a deeper context analysis for determining the actual place it represents.

4. The element is not a geographic location and it was misinterpreted during the recognition process (e.g., *Austin* the personal name or the *Texas* capital). This case corresponds to a non-geo/geo ambiguity. It also requires a deeper context analysis and in the case that it finally results in a location, we should provide its geographic coordinates.

These problems are extremely complicated, mainly considering that we have up to 6 political administrative levels, and that some place names could have several alternatives.

#### The strategy

The TR process returns a list of toponyms $\{t_1, t_2, \ldots, t_n\}$ occurring in the document. Each toponym has associated an extra information (denoted $t_i$.`info`) related to the particular rule that matched it (see Table 4.1). In practice, a toponym that matches the rule number III contains as extra information the place name that follows it in the text, while a toponym matching the rule number II, V or VI records the `feature_class` related to it.

A first request to GeoNW allows to gather all possible alternatives for $t_i$. An element alternatives are denoted $t_i[a_{i1}, a_{i2}, \ldots, a_{ik}]$. This is the input for the disambiguation algorithm, which is a 7 step process. These steps are summarized below:

1. Filtering alternatives according to the extra information

2. Assigning to each alternative the cosine distance from its description to the document.
3. Computing the relational coefficient of each alternative and its hierarchical neighbors at document level.
4. Computing the relational coefficient of each alternative and its hierarchical neighbors at sentence level (see definition 4.1.1).
5. Assigning to each alternative a score resulting from the multiplication of the cosine distance and the relational coefficients at both, document and position levels.
6. Removing elements with low confidence.
7. Assigning the winner alternative to each toponym.

The first step aims to reduce the searching space using the extra information obtained from the recognition process. In this way, alternatives that do not satisfy one of the following conditions are removed from the alternatives list.

- if the extra information is the `feature_class`, then remove all the alternatives not belonging to this class. This condition is applied when the toponym name contains its `feature_class` name. Intuitively, the `feature_class` that should correspond to the toponym is the one that is contained in its name. For example, *Library of Congress* corresponds to the `feature_class LIBRARY`. Thus, there is no sense to consider any other alternatives with a different `feature_class`.
- if the extra information is a place name, then remove all the alternatives that do not contain this place name in their description or are more than 200 km far away. The former is also very intuitive, Alternatives hierarchically related to the place name have a much higher probability of being the actual geographic reference. The latter is considered because there could be places that are close physically but not according to the hierarchy, for example, the cities of each side of an international border.
- if no elements remain after the filtering process, then discard the filtering and reconsider all the original alternatives. Some times a set of elements can be enumerated and be related in other ways, for example, they are all capitals from different places in the world. Because of that, it is not a desire to let the above conditions determine if a possible toponym is a location or not.

The second step uses the toponyms descriptions present in GeoNW. The analysis is similar to the one made in the integration process of Section 3.2.4. The distance between the description of each alternative and the document is computed using the standard cosine similarity function. This distance is always greater than 0 because at least the toponym name appears in the text. It is worth to notice that elements including information from WordNet have more and clearer descriptions.

The third and fourth steps are based on the below definition.

**Definition 4.1.1** (Relational Coefficient)**.** Let $a_i \in GeoNW$ be an alternative of the toponym $t$, let $d$ be the cosine distance between $a_i$ description and the document computed in the previous step and let *toponym_set* be the set of toponyms extracted from an unstructured text. The relational coefficient is defined as:

$$rel\_coef(a_i, toponym\_set) = \begin{cases} 1.4 * d & \text{if } \exists x \in names^*(a_i) : x \in toponym\_set \\ 1.2 * d & \text{if } \exists x \in parent^*(a_i) : x \in toponym\_set \\ 1.2 * d & \text{if } \exists x \in children(a_i) : x \in toponym\_set \\ 1.2 * d & \text{if } \exists x \in siblings(a_i) : x \in toponym\_set \\ 1.1 * d & \text{if } \exists x \in path(a_i) : x \in toponym\_set \\ 1 * d & \text{otherwise} \end{cases}$$

The relational coefficient can be computed for any piece of text. It means that it can be used at document level, at paragraph level, at sentence level or even in a neighborhood containing the toponym $t$. In this proposal it is used in the third and fourth steps, at document and sentence level respectively. The idea behind this coefficient is to increase the confidence of an alternative if there are other elements in the text that are related to it. The increasing values are 40%, 20% or 10% according to the type of relationship.

The fifth step is intended to combine all the values obtained in the last three steps. To each alternative is assigning a score that is the product of the cosine distance and both relational coefficient values.

The sixth step assigns to each toponym in the document the alternative with the highest score, solving the ambiguity problem.

Finally, the seventh step removes those toponyms with low confidence:

- a toponym detected in the recognition process with a low probability of being a location and with a relational coefficient at document level equal to 1 is discarded. This condition dismisses less probable locations that appear only once in the document.
- The average score *avs* of all the winner alternatives, corresponding to toponyms classified with high confidence by the recognition algorithm, can be considered an indicator of how much related are the geographic elements in the document among themselves. Then, winner alternatives of toponyms with low confidence with a score less than $1.25 \times avs$ are also discarded.

## 4.2 Geographic Footprint

Geographic footprint refers to how an unstructured text is represented according to their geographic information. It aims to summarize all the relevant geographical data extracted during the toponym resolution process (see Section 2.3). To accomplish this goal, we have to face two main tasks:

1. definition of a weighting strategy for quantifying the influence of each geographic element in the text.

2. definition of the structure that represents the geographic information in the text.

Like in standard IR systems, the weighting strategy is intended to evaluate how relevant are the geographic elements present in the text. It takes into account the information obtained from the TD process (see Section 4.1.2) and assigns to each geographic element a value that represents its relevance in the text. It is worth to notice that this process is different in queries and documents, because the TD algorithm used for queries is different from the one used for documents (see Section 4.2.2). The main reason of this distinction is that queries are usually much shorter than documents, containing a small number of toponyms (usually one or two) [91]. Documents are richer and mostly contain enough information for the disambiguation. The following two paragraphs describe each of these strategies highlighting the main idea behind their design.

### 4.2.1 Geographic Weighting Strategy in Documents

The weighting strategy here proposed attempts to favor those toponyms in the document, that have high frequency and are best related to other toponyms. Frequency is a very straightforward feature that has been widely used as an indicator of relevance. In the particular case of geographic elements, a toponym which is continuously mentioned in a document should increase its importance with respect to one referred only a few times. Unfortunately, the frequency is not good enough for determining how relevant is a topoynm. Imagine a document where its toponyms have been equally mentioned. Are actually all the toponyms equally relevant? Which toponyms should be more relevant? Due to this lack of accuracy, a deeper analysis must be done. In this sense, the relationships among the toponyms in a document can be very effective features. In this work, we use two different types of relationships: hierarchical and physical. The former refers to the political-administrative hierarchy of the world, while the latter is based on the geographical distance between toponyms. According to these relationships, we define groups of toponyms that are hierarchically or physically related. It is worth to notice that these groups are not exclusive. It means that two toponyms can be related because they are physically close or because they are connected through the political-administrative hierarchy. A formal definition of both, the hierarchical and physical group as well as the necessary concepts related to them can be found below.

**Definition 4.2.1** (Hierarchical Group)**.** Let $G = \{t_1, t_2, \ldots, t_n\}$ be a set of disambiguated toponyms. $G$ is a hierarchical group iif:

$$\forall\ t_i, t_j\ \in\ G,\ \exists t_k\ \in\ G\ :\ t_k \in \{path(t_i) \cap path(t_j)\}$$

**Definition 4.2.2** (Document Geo-radius)**.** Let $G = t_1, t_2, \ldots, t_n$ be a set of disambiguated toponyms belonging to a document $d$. The *geo_radius* of $d$ is defined as:

$$geo\_radius(d) = \frac{\sum_{i=1}^{n} distance(t_i, midpoint(G))}{n}$$

where $midpoing(G)$ refers to the geographic midpoint[2] corresponding to all the locations in $G$.

[2] http://www.geomidpoint.com/calculation.html

**Definition 4.2.3** (Physical Group). Let $G = \{t_1, t_2, \ldots, t_n\}$ be a set of disambiguated toponyms in a document $d$. $G$ is a physical group iif:

$$\forall\ t_i, t_j\ \in\ G,\ distance(t_i, t_j) < geo\_radius(d)$$

**Definition 4.2.4** (Geographic Frequency). Let $t$ be a disambiguated toponym in a unstructured document $d$. The geographic frequency of $t$ in $d$ is defined as:

$$geo\_freq(t, d) = \sum freq(x) : x \in names^*(t)$$

where $freq(x)$ is the number of occurrences of the element $x$ in the document.

**Definition 4.2.5** (Toponym Weight). Let $G$ be a hierarchical or physical group and let $t.dis\_score$ be the score assigned to $t$ by the TD algorithm (see Section 4.1.2). The weight of each toponym $t \in G$ is computed as:

$$w(t, G) = geo\_freq(t, d) * t.dis\_score * |G|$$

where $|G|$ is the cardinality of $G$.

For each toponym we can now obtain two weight values $t.h\_weight$ and $t.p\_weight$. The former corresponds to the weight associated to the toponym after grouping all toponyms in a document according to their hierarchical relationship. The latter corresponds instead, to the weights given to each toponym after grouping by the physical relationship. Notice that by definition the distinct values between both weights is given by the cardinality of their group (see Definition 4.2.5). This fact denotes that the distinct weight values are actually determined by the corresponding relationship.

Finally, the weight of a toponym in a document is defined as the average of both weights, $t.h\_weight$ and $t.p\_weight$. This combination is valid because, by definition, values assigned in the hierarchical groups are comparable with values assigned to the physical groups.

**Definition 4.2.6** (Document Geographic Weighting). Let $t$ be a disambiguated toponym present in a document $d$, let $toponym\_set$ be the set of all disambiguated toponyms in $d$:

$$w(t) = \frac{t.h\_weight + t.p\_weight}{2}$$

Theoretically, the hierarchical group has been designed to include toponyms that are related at different political-administrative levels. For example, a set of toponyms like *Europe, Italy, Rome* and *Vatican City* is a hierarchical group. The physical group, instead, should include toponyms that are not necessarily involved in the political-administrative hierarchy. For example, a set of toponyms like *Tel-Aviv, Valleta, Antalya, Dubrovnik* and *Mediterranean Sea* is a physical group.

We notice that the main feature of the physical group is its ability to scale according to the distance range used in the groups of elements in the document. This is shown in the partitions generated by this group criteria. For example, consider a document referring to *Madrid, Rome, Berlin, Tokyo* and *Pekin.* In this case, we would obtain two groups, one corresponding to the European cities and the other to the Asian ones. There are not references in the document to any of the ancestors of these cities, though.

The purpose of the weighting strategy is then, to find an equilibrium between these two kinds of relationships when favoring “best related” elements in the document. Therefore, the final weight is computed as the average of the weights obtained by both groupings. As a direct effect, the elements that are best related, from both the hierarchical and the physical point of view, are the ones that get the highest weights.

### Weighting implicit geographic references

So far, we have described how to quantify the influence of each geographic reference explicitly mentioned in the document. However, we consider that a document implicitly refers also to those locations containing the ones explicitly mentioned. For example, a document mentioning a number of cities in a country is implicitly referring to the country itself. We therefore, proceed to discuss a weighting strategy for these implicit toponyms.

**Definition 4.2.7** (Implicit reference)**.** Let $t' \in GeoNW$ be a toponym, let *toponym_set* be the set of all disambiguated toponyms in a document $d$. The toponym $t$ is implicitly referenced in $d$ iif:

$$t' \notin toponym\_set \wedge \exists\ t \in toponym\_set\ \ : t' \in path^*(t)$$

**Definition 4.2.8** (Path weight)**.** Let $t$ be a toponym in a document $d$, let *toponym_set* be the set of all disambiguated toponyms present in $d$. We define the weight of the $path^*(t)$ as:

$$w_{path}(path^*(t)) = \min_{t_i \in path^*(t) \wedge t_i \in toponym\_set} w(t_i)$$

**Definition 4.2.9** (Implicit reference weight)**.** Let $t' \in GeoNW$ be an implicit reference of a document $d$, let *toponym_set* be the set of all disambiguated toponyms present a document $d$. Let $toponym\_set|_{t'}$ be the set of elements $t_i \in toponym\_set$ such that $t' \in path^*(t_i)$. The weight of $t'$ in $d$ is defined as:

$$w_{imp}(t') = \max_{t_i \in toponym\_set|_{t'}} w_{path}(path^*(t_i))$$

The Definition 4.2.7 formally defines what is an implicit reference. In a single path, the weight of an implicit reference is equal to the weight of the path (see Definition 4.2.8). However, notice that, such implicit reference may belong to the paths of more than one explicit reference. In this case, the weight of the implicit reference is chosen as the maximum weight it gets among all the paths it belongs to (see Definition 4.2.9).

Selecting the minimum weight value among all explicit toponyms in a path, ensures that implicit references will not get a weight greater than those of explicit ones. Choosing the maximum among all the possible weights of an implicit reference assigns to it a value according to the relevance of the toponyms implicitly referring to it.

### 4.2.2 Geographic Weighting in Queries

In general, a query is much shorter than a document and subsequently, it usually contains only a few toponyms. While recognition process presented in Section 4.1.1 can be easily used, the proposed disambiguation algorithm (see Section 4.1.2) is not compatible with queries size restriction. TD algorithm is mainly based on the relationships among toponyms. Thus, using this technique in queries would produce undesired results. For example, let us think in a query with only one geographic reference, which is a very common scenario. Using the proposed TD algorithm, the second, third and fourth steps cannot be applied because all of them depend on the relationships between the toponym alternatives and the other toponyms in the text. Moreover, there is no sense in applying the fifth step either because of, the already mentioned, queries size restriction. Also, toponyms descriptions include specific information about the toponym itself and usual user queries do not contain this kind of information, thereby there is no much sense to look for similarities between a toponym description and the complete query text.

All these inconveniences and the lack of any other kind of information that can be inferred from the query context rise us to design a sort of semi-disambiguation algorithm. The technique is exactly composed by the first step of the TD algorithm previously proposed. It is called semi-disambiguation because it does not necessarily disambiguate all the toponyms. Remind that the filtering process just helps to reduce the alternatives list of a toponym, removing noticeable wrong geographic locations. After the filtering, all the remaining alternatives are considered. Therefore, during the weighting process a weight value is assigned to each alternative.

The question that arises here is how to properly determine these weight values? At this point of a GIR system process, the document collection has been analyzed and the geographic information present in each document has been extracted and represented. Thus, why not to get some feedback from the document collection geographic features?

For each ambiguous toponym in a query we have the frequency of its alternatives in the whole document collection. This information denotes how popular is an alternative in the whole collection. This allows to think that an alternative with a high frequency in the collection has a high probability of being the one referred by the user query. The following definition formally describes this feature of the collection data.

**Definition 4.2.10** (Toponym Frequency in a collection)**.** Let $t$ be a disambiguated toponym. The frequency of $t$ in a collection of $n$ documents $D = \{d_1, d_2, \ldots, d_n\}$ is defined as:

$$total\_geo\_freq(t) = \sum_{i=1}^{n} geo\_freq(t, d_i);$$

**Definition 4.2.11** (Query Geographic Weighting)**.** Let $\{a_1, a_2, \ldots, a_n\}$ be the alternatives list of a toponym present in a query. The weight of an alternative $a_i$ is defined as:

$$w(a_i) = \frac{total\_geo\_freq(a_i)}{\sum_{j=1}^{n} total\_geo\_freq(a_j)}$$

In summary, the query analysis is composed by a semi-disambiguation algorithm and a weighting strategy based on the probability of an alternative being the one referred by the user. The semi-disambiguation technique allows to consider all the alternatives that remain after the filtering process. This idea came up because there is not enough information about a toponym present in a query. After that, the weighting strategy gives to each alternative a probability value according to the frequency information in the document collection. This weight value can be also interpreted as the relevance of an alternative in the whole collection.

### 4.2.3 Document and query representation

Previously, it was defined the weighting strategies used for documents and queries. The results of both processes are:

- For each document: The list of all the implicit and explicit disambiguated toponyms with their corresponding weights.
- For each query: The list of the remaining alternatives related to each toponym with their corresponding weights.

Thus, a simple geographic footprint can be defined as a set of pairs $(w, t)$, where $t$ is a toponym and $w$ its corresponding weight. The following is a more formal definition of the structure.

**Definition 4.2.12** (Geographic Footprint)**.** Let *text* be a document or a query. Let *toponym_set* be the list obtained by applying the corresponding weighting strategy (see definition 4.2.11) to *text* and let $w_i$ be the weight value assigned to the toponym $t_i \in toponym\_set$. The geographic footprint of *text* is defined as:

$$geo\_footprint(text) = \langle (w_1, t_1), (w_2, t_2), \ldots, (w_n, t_n) \rangle$$

where the pair $(w_i, t_i)$ represents the weight $w_i$ of $t_i$ in *text*

## 4.3 Geographic Ranking Process

Once the whole document collection and the user query have been represented according to their geographic information, the next step is to rank the documents that are relevant for the user query. During the weighting processes the following important analysis has been done in order to improve the accuracy of the geographic ranking.

- Introduction of alternate names in both, document and query analysis: It allows to take into account different ways of naming the same place. Thus, queries that use one alternate name can consider documents that are using another one.
- Definition of two criteria for grouping related toponyms in order to evaluate their relevance in documents: It provides two different points of view regarding the relationships among toponyms, hierarchical and physical. The first one groups toponyms related by their political-administrative relationship and the second one by their geographic distance (see Section 4.2.1).

- Introduction of implicit geographic information in documents: It allows queries consider documents that do not contain the toponym itself but another one closely related. It is worth noting that implicit toponyms in queries are not introduced because toponyms are not completely disambiguated, therefore adding implicit geographic information will introduce a lot of noise.

The result of the above analysis is reflected in the geographic footprint previously defined which allows a very straightforward definition of similarity function.

**Definition 4.3.1** (Similarity Function)**.** Let $\langle (w_1^q, t_1^q), (w_2^q, t_2^q), \ldots, (w_n^q, t_n^q) \rangle$ be the geographic footprint of the query $q$ and let $\langle (w_1^d, t_1^d), (w_2^d, t_2^d), \ldots, (w_m^d, t_m^d) \rangle$ be the geographic footprint of the document $d$. The similarity between a query $q$ and a document $d$ is defined as:

$$geo\_sim(q, d) = \sum_i \sum_k (w_i^q * w_k^d)_{t_i = t_k}$$

**Definition 4.3.2** (Geographic Ranking)**.** Let $geo\_score_i = geo\_sim(q, d_i)$ be the geographic similarity between the query $q$ and a document $d_i$. The geographic ranking list of documents is defined as:

$$geo\_rank(q, D) = \{geo\_score_i\}_{i=1,\ldots,n} : geo\_score_i < geo\_score_j \; \forall i < j$$

Each document in the collection is then evaluated by the above similarity function obtaining the geographic ranking list of relevant documents according to an specific query.

## 4.4 Textual and Geographic Ranking Combination

In this section we describe how the proposed techniques for geographical analysis can be integrated in a full Geographic Information Retrieval search engine.

The whole process can be seen as a model that separately analyses textual and geographic information present in texts, see Figure 1.1. Both components, Textual Analysis and Geographic Analysis, include the Information Extraction, Indexing, Retrieval and Ranking processes. The final output is a ranked list of documents which are relevant to a specific user query (Figure 1.1). This final output is expected to satisfy the user needs better than traditional IR search engine.

In order to produce a combined result, both ranked lists should be merged into one final ranked list. The combination of the textual and geographic rankings constitutes a very sensitive step of the GIR process. For example, the geographic relevance results can be used to re-rank the textual results, or can be combined with the text results at a lower priority [94]. In [74], Martins mentioned different ways of combining text and geographic relevance, such as linear combination of similarity, product of similarity, maximum similarity. In this work we use a linear combination strategy obtained from experimental results.

A first step to integrate the resulting ranked lists is to normalize both the original rankings. The normalization process is even more important for linear combination techniques. There are several normalization approaches for facing this task. In our work we have chosen the **maximum norm**. Namely to divide every entry in the ranking by the maximum raking value in the contained list.

Once both lists have been normalized we can apply the combination strategy. As mentioned before we have chosen a very simplistic linear combination approach. It is inspired in [28].

**Definition 4.4.1** (Ranking Combination)**.** Let *geo_rank* and *text_rank* be the normalized geographic and textual ranking list of the query respectively. Let $geo\_sim(q, d_i)$ and $text\_sim(q, d_i)$ the geographic and textual similarity values between the query $q$ and the document $d_i$. The ranking combination function is defined as:

$$comb\_rank(geo\_rank, text\_rank) = \sum_{i=1}^{k} \beta * (geo\_sim(q, d_i) + text\_sim(q, d_i))$$

where $\beta$ takes value 2 if the document is in both, *geo_rank* and *text_rank*, 1 if the document is only in *text_rank* and 0.5 if the document is only in *geo_rank*.

This strategy benefits documents retrieved in both lists and penalizes those that were retrieved only by their geographic information. It is based on the assumption that documents whose textual information is not relevant to the query will have less importance than those that do.

## 4.5 Conclusions

In this chapter we have described the theoretical approach to solve each one of the tasks that conform a GIR system (as described in 1.1).

For the toponym resolution stage, we divided the problem in two processes: TR and TD. In the first process the geographic references are detected. In the second process a unique location is assigned to each one of the detected geographic references. This task, may be cumbersome due to the fact that many places in the world share the same name.

In the case of the TR problem our solution defines a set of rules that depends on information extracted from the GeoNW resource and on a tagging process that relies on the Stanford CoreNLP [70] tools. The definition of such rules has been made mainly based on an experimental process.

In the case of TD process, we defined a so called *disambiguation score* that is assigned to each possible alternative. This score is calculated as the combination of three values: the cosine similarity between the document and the element description, a relational coefficient among related elements in the rest of the document, and a relational coefficient among related elements in the same sentence the toponym appears. Finally, the alternatives with the highest scores are chosen.

Once the elements in the document have been disambiguated, they contain the information corresponding to a specific location in the world. For the construction of the document footprint, we store for each document the geographic references and assign to each one of these a custom weight. This weight is calculated according to the physical and topological relations between the element and the other elements in the document, the score that was associated to the element in the disambiguation process and the frequency of the toponym (including its alternate names) in the document.

The query processing and representation, follows almost the same process as documents. However, there is one major difference. Often, in queries, there is not enough information for the disambiguation. We notice that the disambiguation process depends on the other geographic references in the context and that queries, in general, contain very few of these [38,91]. To cope with this issue we have designed a sort of semi-disambiguation algorithm that may produce more than one possible location per toponym. The weighting model assigns then, a value quantifying the probability of each one of the possible locations. This value is calculated based on the frequency of the location in the whole document collection.

The design of the document and query footprints make the calculation of the similarity between queries and documents pretty straightforward. Actually, this similarity is calculated as the *dot product* of the vectors containing the document and the query weights.

Finally this chapter concludes with the combination of the document ranked lists resulting from the retrieval process in both the textual analysis and the geographical analysis. The technique followed in this work considers to give higher values to the documents appearing in both resulting rankings, and penalizing documents appearing only in the geographical analysis results under the assumption that the thematic similarity has higher relevance than the spatial one.

# Chapter 5

# Implementation of the GEIR solution

In this chapter we propose a flexible structure for the implementation of a GIR system. In Chapter 1, Figure 1.1 we showed a basic architecture of such systems. The main idea behind this structure is to provide an easy mechanism to test and evaluate new approaches. The structure we propose follows a modular composition making possible to extend or modify each one of the components in a GIR system.

Our implementation, called GEIR: Geographically Enhanced Information Retrieval, is built on top of an existing search engine: Terrier [67], some of the other search engines alternatives are discussed in Section 5.1. As most search engines, Terrier provides a wide set of out-of-the-box plug-ins for standard information retrieval tasks. It covers the textual extraction, the indexing, the retrieval and the ranking processes. It also provides a basic module for the evaluation of an IR system. Moreover Terrier's architecture allows to develop new plug-ins and hook them into the IR process.

As described previously, a GIR system may be divided into two different process, one covering the textual analysis and the other covering the geographical analysis. For the textual process GEIR relies completely on the out-of-the-box tools provided by Terrier. For the geographical analysis GEIR still uses Terrier, but adding the necessary plug-ins for geographic related tasks.

The geographical analysis component of GEIR has been designed also following a modular approach. The main idea is to allow each one of the modules to be interchangeable with other implementations to ease the testing and evaluation of other approaches. As Terrier, GEIR exposes a set of properties that allow to specify the implementations to be used for each one of the GIR tasks. Finally in this chapter we also present the implementation details of the GeoNW knowledge base presented in Chapter 3.

## 5.1 Search engines

A search engine is a tool that allows to index a series of documents and then perform queries against them, providing a, probably sorted, list of documents satisfying the query. There is a wide set of open source search engines available for downloading. Some of these search engines provide a plug-in architecture allowing to substitute components of the tool with custom ad-hoc implementations. In this case, these tools can be also considered search engine frameworks. Usually a search engine framework abstracts from the development of common tasks in IR, like the parsing of documents, the storage of the indexing, the implementation of standard similarity models, etc. For this work we have analyzed some of the existing tools for this purpose. We notice that there is a huge number of similar tools in this area, for a detailed survey on this topic we refer to [75]. Below, we briefly review four of the analyzed tools:

- **Lucene[1]:** is a high-performance, full-featured text search indexing and searching library written entirely in Java. Lucene is highly reputed for its performance and scalability, and is vastly used worldwide. Lucene is developed by the Apache Foundation.

- **Nutch[2]**: is an open-source search engine implemented in Java, which uses Apache Lucene. It is a very efficient search engine, but lacking some state-of-the-art ranking algorithms, such as Okapis BM25. One of its key features is the ability to extend its functionalities through the use of self contained software plug-ins. Nutch is also developed by the Apache Foundation.

- **Lemur Toolkit[3]:** is an open-source toolkit designed to facilitate research in language modeling and information retrieval. Lemur supports a wide range of industrial and research language applications, such as ad-hoc retrieval, site-search, and text mining. Lemur is implemented in C/C++.

- **Terrier[4]:** is a modular platform for the rapid development of large scale IR applications, providing indexing and retrieval functionalities, developed by the Information Retrieval Research Group of the Department of Computing Sciences of the University of Glasgow. Terrier has various cutting edge features, including parameter-free probabilistic retrieval approaches (such as Divergence from Randomness models), automatic query expansion/re-formulation methodologies, and efficient data compression techniques. Terrier is written in Java.

The tool chosen for this work was Terrier. The main reason behind this decision is the modular architecture it provides which makes straightforward its extension. Another important fact was the number of out-of-the-box features already implemented in terrier. We give more details about this in Section 5.2. We notice, however, that among the analyzed tools Lucene is perhaps the most suitable one for a production scenario. We find, indeed, that Terrier has been designed in a way that is more oriented to research tasks.

## 5.2 Terrier role in GEIR

As stated previously, Terrier is the core engine, on top of which we build GEIR. In Chapter 1, Figure 1.1 we showed a generic architecture of GIR systems.

[1] `http://lucene.apache.org/`

[2] `http://lucene.apache.org/nutch/`

[3] `http://www.lemurproject.org/`

[4] `http://terrier.org/`

There were two main components the textual analysis process and the geographic analysis process. In this section we describe how GEIR uses Terrier for implementing both of these processes.

Terrier, as one of the most mature search engine frameworks, provides numerous out-of-the-box features related to common IR tasks [67]. Some of these features, the ones that were key in the selection of Terrier as the core of GEIR are listed below:

- **Open source:** Terrier is open source and very well documented.
- **Modular**: it follows a modular (plug-in based) architecture which allows to substitute existing components with custom ones.
- **Multiple collection format support:** it provides out-of-the-box support for processing the most common formats in existing corpora. For example, *html*, *xml*, *trec* and *warc*; it also supports standard document formats like: *doc*, *pdf* and *txt*.
- **State-of-the-art retrieval approaches:** it provides the implementation of numerous retrieval models, including the DFR[5] model, BM25 and the term dependence proximity models.

The GEIR component for textual analysis completely relies on Terrier. GEIR configuration allows to select among the Terrier alternatives for each one of the elements in the textual analysis. Thus allowing to choose among the different kind of weighting models, weight strategies, etc. This allows, for example, to evaluate the impact of different weighting models for textual analysis into the overall GIR system (see Chapter 6). This component also allows to use the results of standard IR techniques as baseline reference for the evaluation of the GIR system.

On the other hand, for the geographical analysis, GEIR uses Terrier as an underlying framework. The geographical analysis component, as described in previous chapters, provides the mechanisms for geographic aware information retrieval tasks. By default, Terrier, does not provide a geographic indexer, or a geographic weighting model. For this reason custom implementation of all the geographic processing needs to be added. Terrier provides an easy mechanism to plug in new implementations of the IR tasks. The Figure 5.1 shows the schema of the geographic analysis model, the elements in bold black are custom components added to Terrier in order to deal with geographic information.

As displayed in Figure 5.1 the geographic component of GEIR only keeps from Terrier the mechanism for parsing the documents in the collection and the final structure of the index. The following paragraphs describe each one of the custom elements added to extend Terrier to cope with the geographical analysis process.

**GeoNW.** The use of external resources in Terrier for aiding in IR tasks is not uncommon. For example, Terrier relies on external resources for the stemming and stop words removal tasks. GeoNW is the external resource that will be used by the custom implementations of GEIR. In particular the modules in GEIR depending on GeoNW are the Geographic Information Extraction module and the Geographic Query Processing module.

[5] Divergence From Randomness

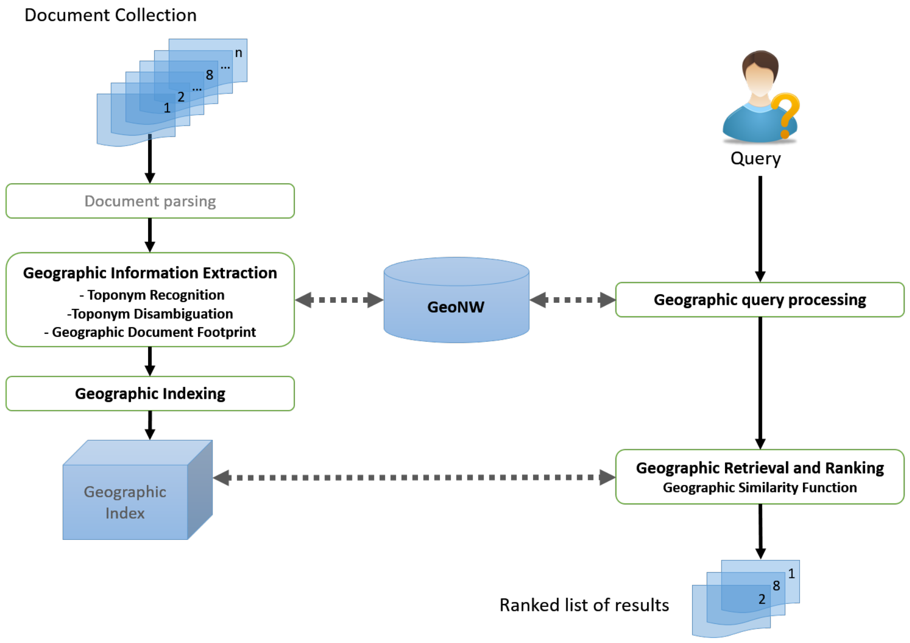

Figure 5.1: GEIR geographic component

**Geographic Information Extraction module.** This module is used just before the indexing process. This is equivalent to the Text Processing module of standard IR solutions. Usually during the text processing stage, operations like the stemming of the text or the elimination of stop words take place. However, in the geographic analysis, this stage takes care of eliminating all non-geographic information. This process is done using first the TR technique described in Section 4.1.1 and then the TD technique described in Section 4.1.2. Also during this stage the Geographic Document Footprint is calculated and every element is assigned with a weight that represents the element relevance in the text (see Section 4.2).

**Geographic Indexing module.** The indexing process will feed the index structure with the information extracted from each document. Since in the previous stage some custom information has been extracted the indexing process should be updated accordingly. For this reason GEIR provides a Geographic Indexer. We remark that GEIR's geographic indexer does not produce a particular index structure. Instead it relies on Terrier's index structures and creates a Direct Index that is afterwards inverted, storing for each geographic term, the list of documents containing it.

**Geographic Query Processing module.** This module is used just before the retrieval process. It takes care of extracting the geographical references from each query. This module implements the technique described in Section 4.2.2. During this stage geographic references in the query are assigned with a weight that represents its relevance

**Geographic Retrieval and Ranking module.** Terrier provides out-of-the-box several similarity measures, however, those are all for calculating text similarity. This module is related to the Geographic Weighting model. It provides Terrier with a way of comparing queries and documents from a geographic point of view. This module produce a final partial result which is the list of documents sorted according to their relevance.

**Ranking combination module.** Finally, one last addition necessary in GEIR is a module for combining the results of the textual analysis component and the geographical analysis component. This module produces the final result to present to the user or to submit to the evaluation process.

## 5.3 GEIR Extensibility

A main goal of the design of the GEIR solution is to keep the modularity inherited from the underlying Terrier framework. The GEIR solution is available at: `https://bitbucket.org/YiyiBB/geirtools`. The GEIR extensions can be divided in two groups. The first one corresponds to the extensions of the indexing functionality. The second, corresponds to the extensions of the retrieval functionality.

### 5.3.1 Extending the indexing functionality

Figure 5.2 shows the GEIR extensibility for the indexing process.

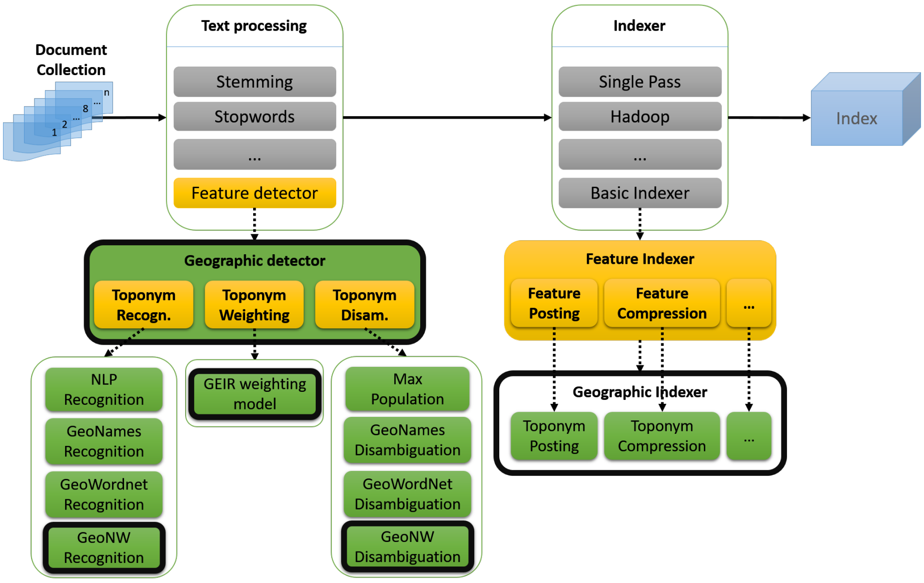

Figure 5.2: GEIR geographic component (indexing)

The figure displays three kind of elements. The elements in gray, corresponds to out-of-the-box alternatives given by Terrier (the three dots are symbolic references to similar elements). The elements in yellow and green correspond to the custom elements belonging to GEIR. Yellow elements represent generic abstract tools, while the green elements symbolize actual implementations of the corresponding abstract functionality. The elements with a dark strong border corresponds to the implementation of the techniques described in Chapter 4.

**Text processing extensions.** The text processing pipeline in Terrier is the set of operations applied to the document text. Usually this processing includes stemming operations, the removal of stop-words and other spurious elements in the text. In the geographic processing, the goal is to drop all non-geographic elements from the text. To this aim GEIR provides a generic *Feature Detector*. This is an abstract module that can be implemented according to the type of features relevant for the analysis. In GEIR the implementation provided is a *Geographic Detector*, however we remark that by extending the *Feature Detector* one can target other kind of features like time expressions, person references, etc.

**The geographic detector extensions.** As stated before GEIR provides a geographic capable feature detector. This module is customizable as well. There are three generic abstract submodules within the *Geographic Detector*. The *Toponym Recognition* module is the entry point for adding new approach for this task. GEIR provides by default several implementations of recognizing approaches. The *Toponym Disambiguation* module allows to implement different disambiguation strategies, GEIR includes a few of these; a naive one, based on the population values, and three more based on different external knowledge bases. The Table 5.1 describes each of the available alternatives in GEIR for toponym recognition and toponym disambiguation. Finally, the *Toponym Weighting* module allows to implement different weighting mechanisms, GEIR includes only one weighting model, the one corresponding to the description in Section 4.2.1.

**Indexing extension.** Terrier provides a wide variety of indexing strategies. Some of them are referred in Figure 5.2. GEIR extends the Terrier's *Basic Indexer* strategy. This strategy follows the classic IR two pass indexing. In the first pass a direct index is created storing all the terms found for each document, in the second pass the index is inverted storing for each term the list of documents that contains it. The indexing module is composed by several submodules that provide the manipulation of the structures that are stored in the index. For example, there is a submodule for describing the terms information (id, weight, list of documents in which it appears, etc.), there is another submodule for compressing, serializing and de-serializing this information from the disk, etc. The implementation of these modules is closely related to the implementation of the text processing modules. GEIR provides the proper implementations matching the accordingly the *Geographic Detector* module.

Table 5.1: GEIR Indexing extension modules

| | **Module** | **Description** |
|---|---|---|
| **Toponym Recognition** | NLP recognition | Performs the recognizing process based on the CoreNLP tools. In practice this implementation relies on the Name Entity Recognition feature, selecting as toponyms those recognized as such by its algorithm. |
| | GeoNames based recognition | An adaptation of the algorithm described in Section 4.1.1 using GeoNames as external knowledge base |
| | GeoWordNet based recognition | An adaptation of the algorithm described in Section 4.1.1 using GeoWordNet as external knowledge base |
| | GeoNW based recognition | The algorithm described in Section 4.1.1 using GeoNW as external knowledge base. This is the default approach in GEIR |
| **Toponym Disambiguation** | Disambiguation based on Max Population | Disambiguates by choosing the alternative with the highest population, this technique has unexpected results, since the population values in the considered knowledge bases are not precise. |
| | GeoNames based disambiguation | An adaptation of the algorithm described in Section 4.1.2 using GeoNames as external knowledge base |
| | GeoWordNet based disambiguation | An adaptation of the algorithm described in Section 4.1.2 using GeoWordNet as external knowledge base |
| | GeoNW based disambiguation | The algorithm described in Section 4.1.2 using GeoNW as external knowledge base. This is the default approach in GEIR |

### 5.3.2 Extending the retrieval functionality

The Figure 5.3 shows the GEIR extensibility for the retrieval process.

As in the previous section, the figure displays three kind of elements. The elements in gray, corresponds to out-of-the-box alternatives given by Terrier (the three dots are symbolic references to similar elements). The elements in yellow and green corresponds to the custom elements belonging to GEIR. Yellow elements represent generic abstract tools, while the green elements symbolize actual implementations of the corresponding abstract functionality. The elements with a dark strong border corresponds to the implementation of the techniques described in Chapter 4.

**Query processing.** As stated before the query processing is very similar to the processing of documents. In fact, GEIR uses the same module for these tasks. The only difference in the GEIR implementation is in the toponym disambiguation process in the queries, see Section 4.2.2.

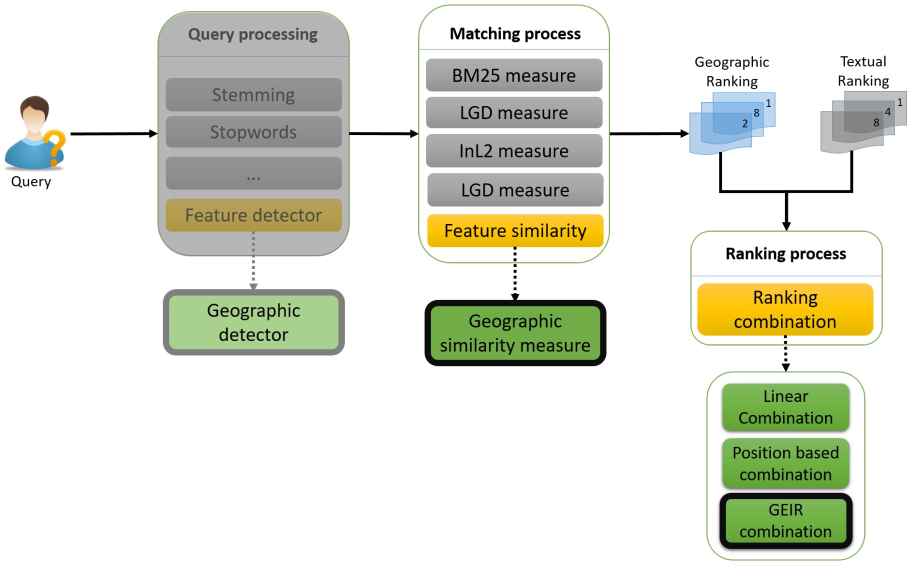

Figure 5.3: GEIR geographic component (retrieval)

**Matching process.** The matching process decides how similar a document is to a particular query. In Terrier several similarity measures are provided out-of-the-box, corresponding to the different weighting models available. GEIR proposes a generic abstract similarity that is the entry point for further similarity measures based on the weighting models used in the *Feature Detector* module. GEIR implements only one similarity measure, the one described in Section 4.3.

**Ranking combination.** The ranking combination is the final task before producing a final result. This task's input are the resulting ranking lists obtained by the textual and the geographical analysis. GEIR provides a generic *Ranking Combination* module with three explicit implementations. The Table 5.2 describe the details of these approaches.

## 5.4 GeoNW Implementation

As described in previous sections GEIR modules often rely on an external geographic knowledge base: GeoNW. The main characteristics of GeoNW have been discussed in Chapter 3. In this section we describe GeoNW from the implementation point of view. GeoNW main source is the GeoNames database. For this reason GeoNW is implemented also as a database. To the information obtained from this database GeoNW adds some extra information obtained from the WordNet ontology. The Figure 5.4 shows the entity relationship diagram of the GeoNW database.

The GeoNW scheme is actually pretty similar to the GeoNames scheme. However the main differences are: (i) the elimination of fields unnecessary in GEIR, (ii) the addition of new tables like *descriptions* and *extra_class* and (iii) the addition of the self reference *linked_to* in table *toponyms*. An important

Table 5.2: GEIR Ranking combinations

| | **Module** | **Description** |
|---|---|---|
| **Ranking combination** | Linear combination | Is the implementation of basic linear combination algorithm the final score for a document is calculated as $\alpha \times text\_score + (1 - \alpha) \times geo\_score$ where $\alpha$ is an argument specified by the user, and *text_score* and *geo_score* are the scores in the textual and the geographic ranking respectively. |
| | Position based combination | A combination based on the position of the document in the textual and geographical rankings. |
| | GEIR combination | In this case the final score for each document is also calculated upon a combination of its textual score and its geographical score. This implementation corresponds to the expression described in Section 4.4 |

characteristic of the GeoNW database is that some tables contain redundant information. This is the case of the tables *hierarchy* and *extra_class*. In both cases the information in these tables can be calculated from the information in the table *toponyms*. However this technical decision has been made attending to two important facts. The first is that GeoNW will not receive update operations, therefore there are not concerns about maintaining the replicated information. The second is that accessing, for example, to the hierarchy information is a rather common operation from within GEIR, therefore the necessity of providing a fast way for calculating this information. The Table 5.3 briefly describes the GeoNW tables. The GeoNW database is available at: `https://bitbucket.org/YiyiBB/girdb`.

### 5.4.1 Efficient access to GeoNW toponyms

Due to the large size of GeoNW it is important to design an effcient way of searching geographic references in unstructured documents. A compressed trie structure [54] is used for this purpose. Trie elements are filled with the names of the GeoNW instances, the trie leaves are filled with the list of all the instances sharing that name. We remark that for this process only centroid toponyms are considered.

Let $c_n$ be the number of characters in a document, $kw$ the number of keywords (geographic references) stored in the trie and $avg$ the average keyword length. A naive approach to check for the presence of all keywords in a document would be $O(c_n * kw * avg)$ which comes from considering each character of the document as the possible start for all keywords.

The use of a compressed trie structure, in particular an Inverted Radix Tree [18], allows to reduce the number of operations to $O(c_n * log(avg))$. Notice that in this way the number of keywords does not affect the efficiency of the scanning process which is key because of the size of `GeoNW` (currently 8 millions of terms approximately).

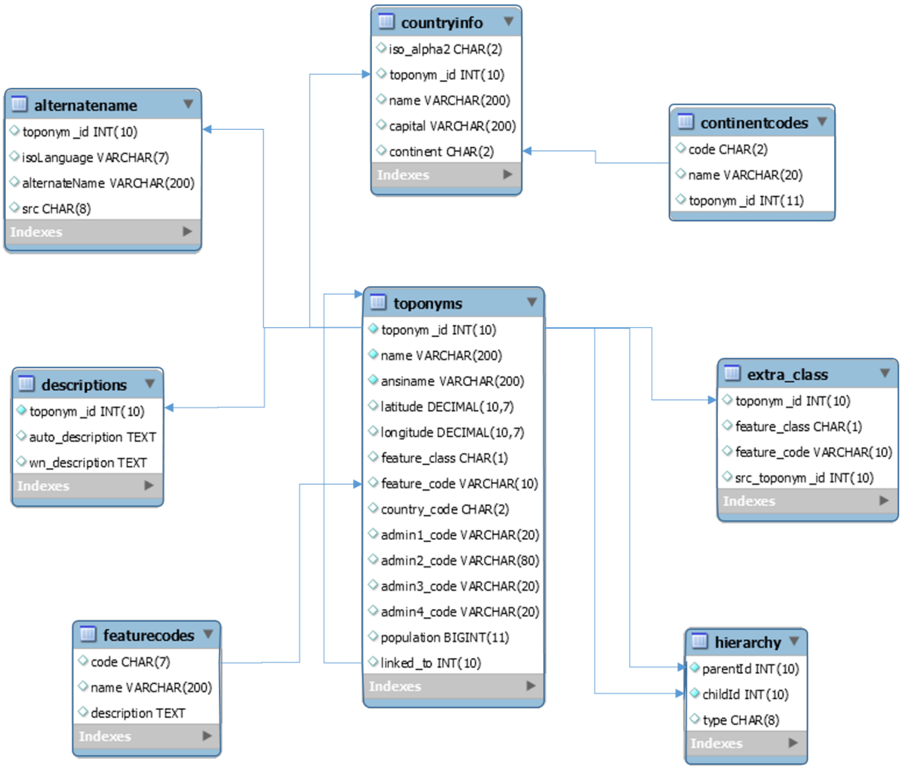

Figure 5.4: Entity relationship diagram of the GeoNW database

## 5.5 Conclusions

In this Chapter we reviewed the design of GEIR, a solution oriented to the analysis of documents paying special attention to the geographic semantics. GEIR is built on top of Terrier, a search engine framework that has been designed mainly for research purposes. These framework has two key advantages a wide set of out-of-the-box tools and a very flexible design that allows, in an easy way, to plug-in new functionality. In this work, we exploit this flexibility by adding a set of plug-ins oriented to the geographical analysis.

An especial interest in GEIR has been put to the fact of producing a flexible solution that allows to plug-in new implementations of classical tasks in the same way its underlying framework does. GEIR contains implementations of all classic tasks in a GIR system. On the other hand, GEIR also includes generic abstract modules that can be extended to add new functionality.

In this Chapter we also reviewed the structure of the GeoNW database, we gave a detailed information on its tables and relationships. Because of the size of the geographical domain, we found necessary to design a strategy for efficiently matching GeoNW's elements in documents. We proposed the use of an Inverted Radix Tree [18] which ensures that the complexity of the detection process does not depend on the size of keywords domain but only in the length of the document.

Table 5.3: GeoNW tables

| **Table** | **Description** |
|---|---|
| *toponyms* | This table is the main one in GeoNW. It contains all the known locations in the world. The elements in this table are imported completely from the GeoNames database removing unnecessary information (e.g., zip code, timezone, elevation data). This table also adds a new field that links toponym elements to other toponyms. The semantic of this *linked_to* relationship is described in Section 3.2.2. |
| *alternatename* | This table contains all the known names for a given toponym. The information in this table is primarily obtained from the alternate names information in GeoNames and then enriched by elements corresponding to the toponym synset in WordNet. |
| *countryinfo* | The information in this table is obtained completely from the corresponding GeoNames table. Its structure is also the same except for elimination of some fields. |
| *continentcodes* | The information in this table is manually inserted. It contains the toponyms representing the world continents. |
| *featurecodes* | The information in this table is obtained completely from the corresponding GeoNames table. Its structure is also the same. |
| *descriptions* | One of the main features of GeoNW is that it contains descriptions of the toponyms present in the database. This description is primarily computed from the implicit information in GeoNames. Afterwards, the descriptions for the elements that are also part of WordNet are enriched. |
| *hierarchy* | This table contains the hierarchical relationships among the toponyms in GeoNW. A similar table already existed in GeoNames, however in that case it contains only non-administrative relationships, since the administrative relationships could be inferred from the information of the toponyms. In GEIR the hierarchy relationship among toponyms is key for several processes, for that reason the *hierarchy* table provides direct access to this information. This table is filled from the information extracted from the toponyms table, and also from the information expressed by the WordNet meronymy relationships. |
| *extra_class* | Another key feature of GeoNW is the linking of elements referring to the same (or close enough) places. When a linking is performed it is possible that two elements of different classes (feature-codes) get linked, see Section 3.2.2. This table allow to relate toponyms representing toponym gropus with more than one class. |

# Chapter 6

# Evaluation

Last chapters describe all the components involved in the geographic analysis process that have been proposed in this work. As mentioned before, we expect to improve ranking results in traditional IR systems by adding this special treatment of the geographic information present in the text. Therefore, this chapter aims to evaluate the actual improvement of our proposal through the use of existing test collections built for this purpose.

The evaluation process analyzes the contribution of each proposed component by comparing them with baselines approaches as well state-of-the-art ones. In summary, next sections include the evaluation of:

- GeoNW: the geographic knowledge base presented in Chapter 3;
- the TR and TD algorithms described in Section 4.1.1 and Section 4.1.2 respectively;
- the geographic weighting strategy explained in Section 4.2.1 and
- the textual and geographic combination ranking approach presented in Section 4.4

The textual analysis, required for evaluating the whole GIR system, was carried out taking the best results obtained from the standard IR models provided by Terrier framework. The best results of Terrier's algorithms were obtained by using $DLH13$ [65] and $LGD$ [30] models. Also we include in the evaluation the $BM25$ approach as it is a well known standard technique in IR applications.

The evaluation was supported by two test collections: TUD-Loc2013 and GeoCLEF test collections. The Table 6.1 shows some of their main characteristics.

Table 6.1: Main characteristics of TUD-Loc2013 and GeoCLEF test collections

| | **GeoCLEF** | **TUD-Loc2013** |
|---|---|---|
| **History** | It was a track oriented to evaluate GIR systems that was organized by the Cross Language Evaluation Forum (CLEF* ) from 2005 to 2008 [69]. There were total of four editions, one for each year. It has become the default standard for evaluating GIR systems, since to the best of our knowledge, up to date there is no other large test collection for this purpose. | It was created at the TU Dresden by Philipp Katz, David Urbansky, and Uliana Andriyeshyna [52]. It emerged as because of the lack of freely available test collections dedicated to evaluate TR and TD algorithms. |
| **Document Collection** | The collection is composed by documents previously used in CLEF campaigns. There are 169,477 documents extracted from the British newspaper The Glasgow Herald (1995) and the American newspaper The Los Angeles Times (1994). The collection contains stories about events occurred around the world, thus it represents a wide variety of geographic places. | The dataset consists of 152 English text documents retrieved from different URLs. The documents mainly contain text that was manually extracted removing elements such as banners, navigation menus, comments, headers, and footers. |
| **Topics** | There were generated a total of 100 topics (25 per year). Most of them can be easily separate in a thematic part and a geographic constraint (i.e. `Trade Unions in Europe`). | Topics are NOT included because the purpose here is only to evaluate the annotation of geographic references. |
| **Judgments** | Judgments were made manually. A pooling approach was used, merging the top ranked documents for each query which were then judged by humans. The average number of documents analyzed per query was 626. The result was stored in a CSV file containing (query number, document ID, corresponding relevant or non-relevant classification) | The annotation of each document was done manually in XML style. Relevant parts of the text were tagged denoting the appropriate types. There was also generated a CSV file containing (document ID, toponym offset, geographic coordinates, source ID*). The whole collection contains 3814 annotations including 55 references corresponding to street names, 37 corresponding to street numbers and 17 corresponding to zip codes. It is worth noting that at the moment our system does not recognize this type of geographic references. |
| **Evaluation Metric** | $MAP$ (see Section 2.6.2) | $F_1$ (see Section 2.6.2) |

* http://www.clef-initiative.eu/web/clef-initiative/home

* The source ID refers to the toponym ID in the external resource that was used. Almost all the toponyms are included in GeoNames knowledge base and subsequently they contain a geonameID

The Figure 6.1 and Figure 6.2 show the place names distribution of both test collections. As it can be seen, although GeoCLEF includes place around the world, higher concentrations of toponyms are located over Glasgow and Los Angeles (where the newspapers are published). On the other hand, TUD-Loc2013 has a wider coverage being Europe, Central America and the United States the areas containing more toponyms in the collection.

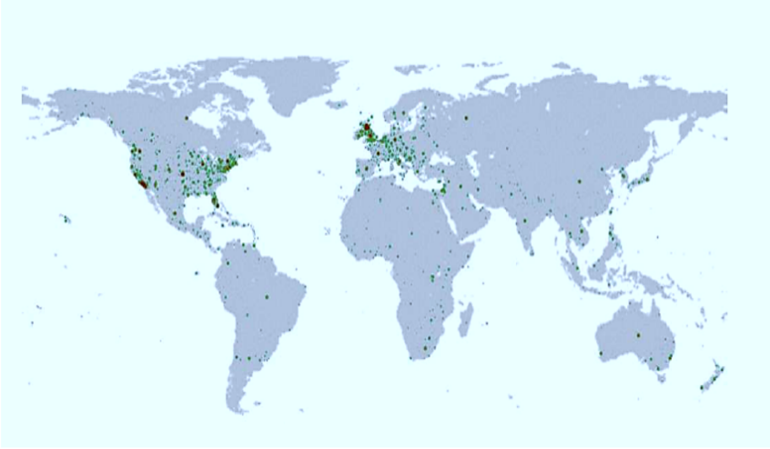

Figure 6.1: Place names distribution in GeoCLEF document collection (image taken from [81])

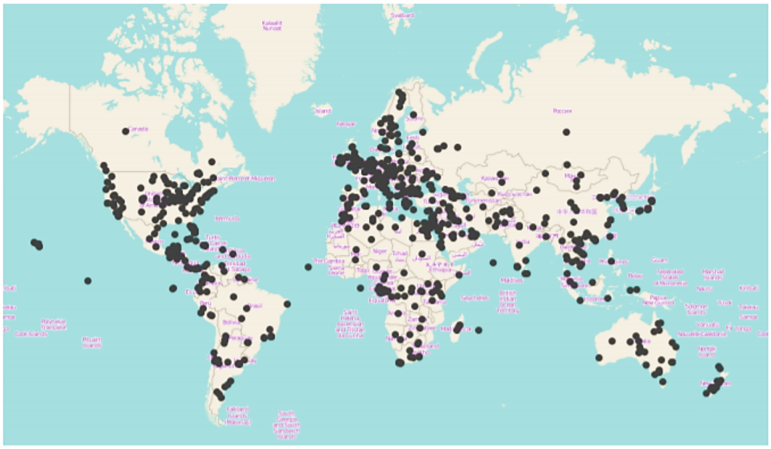

Figure 6.2: Place names distribution in TUD-Loc2013 document collection (image taken from [52])

## 6.1 Evaluation of Toponym Resolution algorithms

The evaluation of the proposed TR and TD techniques was carried out using both test collections: TUD-Loc2013 and GeoCLEF. The former provides an explicit evaluation of the algorithms allowing to compare with other existing approaches. The latter assesses the proposed GIR solution providing an implicit evaluation of the techniques.

### 6.1.1 Evaluation using TUD-Loc2013

TUD-Loc2013 test collection allows to evaluate both techniques: TR and TD. The experiment consists of comparing our proposals with the following approaches:

- Maximum population based approach (`max_pop`): Ambiguous place names are disambiguated taking the geographic location with the highest population.

- Heuristic TD approach (`heuritic`): It was proposed by Katz [52]. It is based on a set of heuristics that use the feature type of the candidate toponym (i.e., `continent, country`), the population number and the geographic distance among other toponyms in the text.

- Machine Learning TD approach (`ML`): It was also proposed by Katz [52]. It use a classifier that was trained using TUD-Loc2013.

- Yahoo! Boss Geoservices[1] (`yahoo`):It is a Web service API that recognizes toponyms in unstructured text. It also provides the geographic coordinates of each recognized toponym.

- GeoNW based TR approach (`GeoNW_TR`): It is the TR algorithm proposed in this work.

- GeoNW based TD approach (`GeoNW_TD`): It is the TD algorithm proposed in this work.

The Figure 6.3 shows the results of the proposed TR algorithm explained in Section 4.1.1, named `GeoNW_TR` and the four different techniques briefly described above. Notice that the recognition process only takes into account the identification of the keywords related to geographic references. This means that the interest in this case is to evaluate the ability of the proposal to determine if a word or sequence of words is a geographic reference or not. The approaches used for comparing with `GeoNW_TR`, are considered TD strategies because all of them return for each toponym in the text a unique geographic location. Therefore, to achieve this experiment there was only considered if the strategy identifies as geographic references the keywords that actually do are. For example, if *Paris* appears in the text and it is a location, we do not care about which *Paris* refers the toponym, we just expect that the algorithm recognizes it as a geographic reference.

[1] https://developer.yahoo.com/boss/geo/

In this case, the *precision* and *recall* measures respond to the questions of which portion of the total recognized keywords were actually geographic references and which portion of the geographic references in the text were properly recognized. As it can be seen in the graphic (Figure 6.3), `GeoNW_TR` achieved a recall of 89.9%. This is the highest recall value among all the other approaches improving the result in more than 10%. The precision was not as good being the `ML` and the `heuristic` approaches about a 10% better. The high recall allows to obtain the highest $F_1$ value among all the other approaches. This result verifies the assumption of the algorithm that aims to identify as much as possible geographic references.

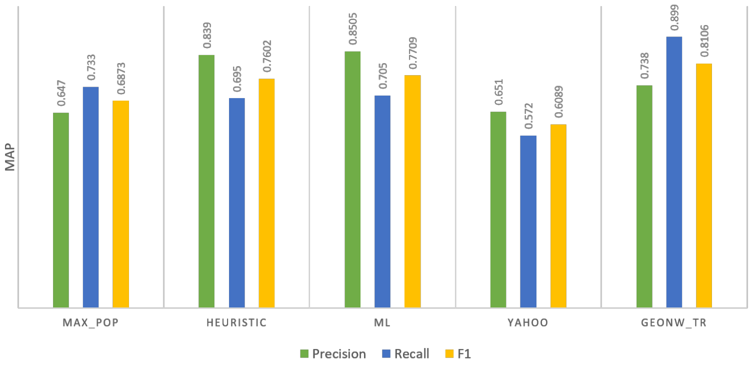

Figure 6.3: Evaluating toponym recognition process using TUD-Loc-2013 test collection

On the other hand, the results of the TD strategy is shown in Figure 6.4. The experiment investigates how well the proposed TD technique, `Geo_TD` (see Section 4.1.2) correctly assign to each geographic term previously identified its corresponding geographical coordinates. The comparison was done with the same four approaches used in the above TR experiment, but in this case the proper identification of the exact location was taken into account.

The *precision* and *recall* measures are intended to evaluate which portion of the locations identified by the system were properly identified and which portion of the geographic locations in the text were properly identified by the strategy. The results of `GeoNW_TD` have a behavior similar to the `GeoNW_TR` technique. The algorithm achieved the best $F_1$ value because of a high recall, which improved the recall of the other approaches in about 20%.

The downside of both, `GeoNW_TR` and `GeoNW_TD` techniques is the slightly low precision value. One reason for this result is the use of a very large geographic knowledge base, which generates a lot of ambiguous place names, but the use of a smaller one can incur in the lost of important information. Notice that these ambiguities come from the real world where there are several places that share the same name. Removing information from the geographic knowledge base could improve this particular result but could affect furthers, depending on the test collection that is used. Currently, attempts to improve the precision are producing a significant reduction of the recall, negatively affecting the F1 measure and the final ranking results. However the current proposal is not as precise as the other two approaches, it can be appreciated from the experiments that our main purpose of improving the ranking results is achieved. Nevertheless, future work has been oriented to improve these precision values.

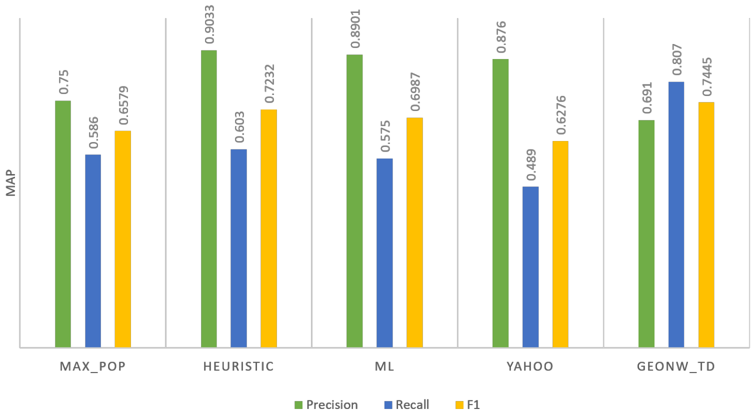

Figure 6.4: Evaluating toponym disambiguation process using TUD-Loc-2013 test collection

### 6.1.2 Evauation using GeoCLEF

The toponym resolution process can be also evaluated using GeoCLEF, the aim is to analyze the behavior of the entire GIR process using the proposed *GeoNW_TD* technique and the *max_pop* approach mentioned above. For the textual analysis were selected the best three weighting models among all weighting strategies implemented on Terrier. The following list describes all the different combinations of the operations in `Text_Proc` and `Geo_Proc`:

- `BM25`: It is a well-known weighting model strategy for textual analysis. The implementation used in this work is included in Terrier framework 4.1. In this case, **the geographic analysis is not included**.

- `LGD`: This model is based on the log-logistic distribution to derive a simplified DFR model [30]. The implementation used in this work is included in Terrier framework 4.1. In this case, **`the geographic analysis is not included`**.

- `DLH13`: This weighting model is a generalization of the parameter-free hyper-geometric DFR model in a binomial case [65]. The implementation used in this work is included in Terrier framework 4.1. In this case, **`the geographic analysis is not included`**.

- `BM25 + GeoNW_TD`: This includes both, the textual and geographic analysis. It is a combination of the BM25 weighting model and our geographic information analysis **including our TD technique**.

- `LGD + GeoNW_TD`: This includes both, the textual and geographic analysis. It is a combination of the LGD weighting model and our geographic information analysis **including our TD technique**.

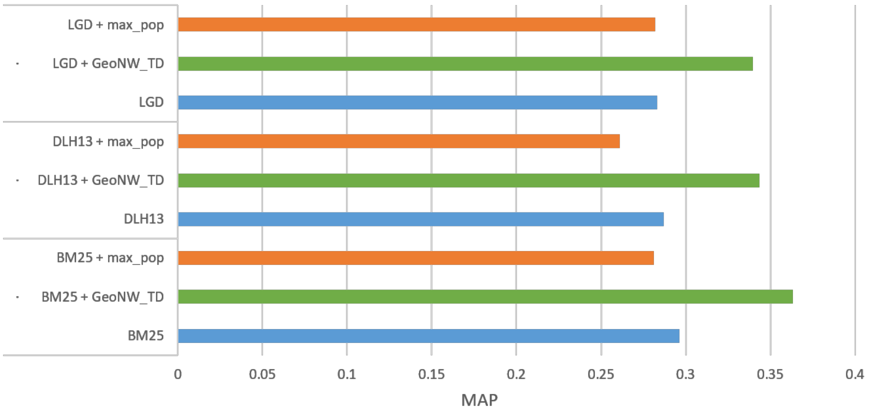

Figure 6.5: `GeoNW_TD` strategy vs maximum population based approach using GeoCLEF test collection

- `DLH13 + GeoNW_TD`: This includes both, the textual and geographic analysis. It is a combination of the DLH13 weighting model and our geographic information analysis **including our TD technique**.

- `BM25 + max_pop`: This includes both, the textual and geographic analysis. It is a combination of the BM25 weighting model and our geographic information analysis applying the **maximum population** based approach.

- `LGD + max_pop`: This includes both, the textual and geographic analysis. It is a combination of the LGD weighting model and our geographic information analysis applying the **maximum population** based approach.

- `DLH13 + max_pop`: This includes both, the textual and geographic analysis. It is a combination of the DLH13 weighting model and our geographic information analysis our geographic information analysis applying the **maximum population** based approach.

The experiment is based on the 100 queries introduced in the four GeoCLEF campaigns. Each result corresponds to the average $MAP$ value of the four GeoCLEF tracks. The Figure 6.5 clearly show how the use of the `GeoNW_TD` technique improves the final ranking results in all the three combinations demonstrating the effectiveness of a proper geographic analysis in unstructured texts. Notice how the maximum population based approach produces a negative effect over the ranking results. This occurs because naive approach does not take into account the context where the toponym appears. Choosing in all the cases the location with the highest population completely removes the possibility to other alternatives of being the one mentioned. For example, $Paris, France$ has a higher population than $Paris, Texas$, thus according to the $max\_pop$ approach the latter never is taken.

## 6.2 Evaluation of GeoNW

In an attempt to evaluate the effectiveness of the proposed geographic knowledge base GeoNW we compare the difference of the Average Precision of each GeoCLEF track (25 queries each one) when using pure GeoNames, GeoWordNet [42] or our GeoNW as the external knowledge source.

The toponym resolution and the relevance processes strongly depends on the features provided by GeoNW. In order to carry out this experiment the needed features were generated according to the information given by GeoNames and GeoWordNet. This information is not exactly the same present in GeoNW because the building process of GeoNW introduces new relationships among the geographic references that are important aspects for our geographic analysis (see Chapter 3).

For each GeoCLEF campaign the proposed GIR system was evaluated using the different combinations of the textual approaches (BM25, DLH13 and LGD) with the geographic ones. In this case, the different geographic approaches were obtained by the variation of the three geographic knowledge bases. The following are the different configurations used in this experiment.

- **BM25_GeoNames**: BM25 + Geographic Analysis using GeoNames.
- **BM25_GeoWordNet**: BM25 + Geographic Analysis using GeoWordNet.
- **BM25_GeoNW**: BM25 + Geographic Analysis using GeoNW.
- **DLH13_GeoNames**: DLH13 + Geographic Analysis using GeoNames.
- **DLH13_GeoWordNet**: DLH13 + Geographic Analysis using GeoWordNet.
- **DLH13_GeoNW**: DLH13 + Geographic Analysis using GeoNW.
- **LGD_GeoNames**: LGD + Geographic Analysis using GeoNames.
- **LGD_GeoWordNet**: LGD + Geographic Analysis using GeoWordNet.
- **LGD_GeoNW**: LGD + Geographic Analysis using GeoNW.

The Figure 6.6 shows the best results obtained from the above configurations. The graphic corresponds to the average precision per query of the best combination that use GeoNames, GeoWordNet and GeoNW. It can be seen how GeoNW outperforms, at least slightly, the results of both GeoNames and GeoWordnet for all the 100 queries. The average improvement of GeoNW over GeoNames and GeoWordNet is about 10% and 5% respectively.

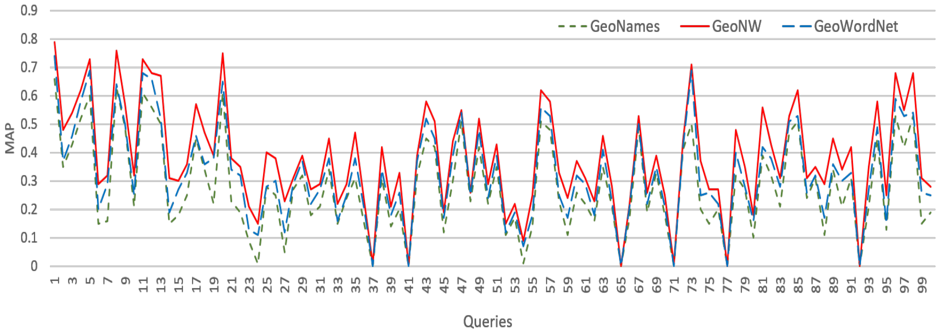

Figure 6.6: GeoNW vs GeoNames using all GeoCLEF campaigns

On the other hand, Table 6.2 shows the MAP values obtained using all four GeoCLEF campaigns separately. Again these are the best results obtained from the different configurations above described. The table clearly reflects the effectiveness of GeoNW which improves the MAP values for all four GeoCLEF campaigns.

Table 6.2: Comparison of GeoNW vs GeoNames vs GeoWordNet using GeoCLEF test collection and MAP measure

| **GeoCLEF Track** | **MAP using GeoNames** | **MAP using GeoWordNet** | **MAP using GeoNW** |
|---|---|---|---|
| 2005 | 0.3547 | 0.4162 | **0.4896** |
| 2006 | 0.2444 | 0.2853 | **0.3252** |
| 2007 | 0.2372 | 0.2801 | **0.3204** |
| 2008 | 0.2768 | 0.3200 | **0.3672** |

## 6.3 Evaluation of the geographic footprint representation model

The aim of this experiment is to measure the effectiveness of the proposed geographic weighting strategy. Remind that the technique is composed by two main relationships among toponyms: i) hierarchical and ii) physical. The former corresponds to the political-administrative division while the latter refers to the geographic distance between locations (see Section 4.2.1). The experiment consists of analyzing each relationship separately in order to verify if the best alternative is actually their combination. Therefore, we have three possibilities:

- **Hierarchical**: Only the hierarchical relationship is considered.

- **Physical**: Only the geographical distance among toponyms is considered.

- **Hierarchical + Physical**: It is the proposed strategy. Both relationships are considered.

Figure 6.7 clearly shows the desired result. The combination of both features reaches the best result increasing the MAP value in more than 10%.

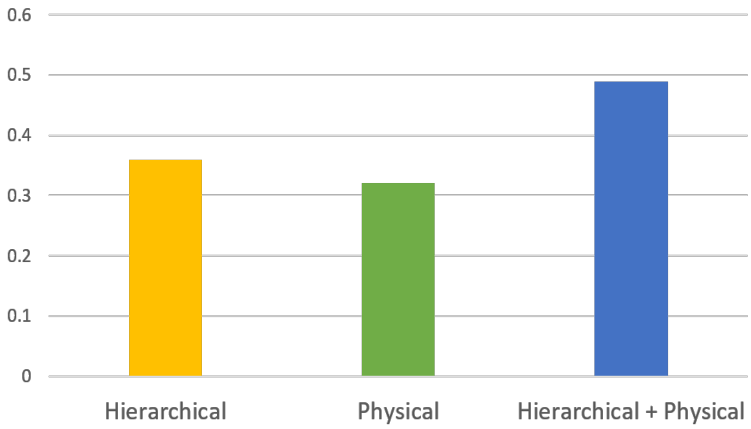

Figure 6.7: Effectiveness of combining topological and geographical distance features

## 6.4 Evaluation of the Combination Ranking Function

The last step in the GIR process proposed in this work is the combination of the ranked lists resulting of the textual and geographic analysis. This experiment is intended to compare different combination ranking approaches in order to analyze their effect over the final result that should satisfy the user needs.

The following are the definitions of the combination ranking techniques used in the evaluation process. The different strategies were previously considered by Martins in [73].

**Definition 6.4.1** (Classic Linear Combination)**.** Let $_T$ and $S_G$ be the textual and geographic score of a document respectively. The classic linear combination strategy is defined as:

$$lin\_comb(S_T, S_G) = \alpha \times S_T + (1 - \alpha) \times S_G$$

where $\alpha \in [0, 1]$ is a parameter thar refe reflects the influence of the textual and geographic ranking over the final ranking list.

**Definition 6.4.2** (Product Combination)**.** Let $_T$ and $S_G$ be the textual and geographic score of a document respectively. The product combination strategy is defined as:

$$prod\_comb(S_T, S_G) = S_T \times S_G$$

**Definition 6.4.3** (Maximum Combination)**.** Let $_T$ and $S_G$ be the textual and geographic score of a document respectively. The maximum combination strategy is defined as:

$$mac\_comb(S_T, S_G) = \max S_T, S_G$$

**Definition 6.4.4** (Step Linear Combination)**.** Let $_T$ and $S_G$ be the textual and geographic score of a document respectively.. The step linear combination strategy is defined as:

$$step\_comb(S_T, S_G) = S_T \times H(S_G)$$

where

$$H(S_G) = \begin{cases} 1 & \text{if } S_G > thereshold \\ 0 & \text{otherwise} \end{cases}$$

The step linear combination is the same ranking according to textual analysis but only geographical relevant documents are considered. The key aspect of this strategy is the selection of the threshold. We tested using as threshold the average geographic score.

**Definition 6.4.5** (CombMNZ)**.** Let $_T$ and $S_G$ be the textual and geographic score of a document respectively. Let *numb_sim* the number of non-zero similarities. The *combMNZ* strategy is defined as:

$$combMNZ(S_T, S_G) = (S_T + S_G) \times numb\_sim$$

The CombMNZ technique, proposed by Lee in [59], provides higher scores to documents retrieved by both the textual and the geographic analysis. Notice that in this particular case, *numb_sim* takes value 2 if the document is ranked in both lists and 1 if it is ranked in one of both. The strategy proposed in this work is based on this technique. The difference between them relies on the value of *numb_sim*. Our assumption is that documents retrieved only by the textual analysis are more important than documents retrieved only by the geographic process.

The experiment compares all the above defined ranking combination with the one proposed in this work. It was also considered using for the textual process the three different weighting model approaches that have been used along with the evaluation process (BM25, DLH13 and LGD). In this way, each combination strategy was evaluated using the following configurations:

- **BM25_Geo**: BM25 + Geographic Analysis
- **DLH13_Geo**: DLH13 + Geographic Analysis
- **LGD_Geo**: LGD + Geographic Analysis

The evaluation was made using the four GeoCLEF tracks. Using the classic linear combination, the best result was achieved with $\alpha = 0.73$ using `LGD_Geo` configuration and the GeoCLEF 2005 track (Figure **??**). As it can be seen in the figure the best $MAP$ value reached was 0.419 which slightly improves the baseline approach using the DLH13 weighting strategy. However, our proposal reports better results using any of the three configurations stated above. In the figure, it is shown the best result obtained using GeoCLEF 2005 track. In this scenario, our combination ranking strategy, using `LGD_Geo` obtains a $MAP$ value of 0.489.

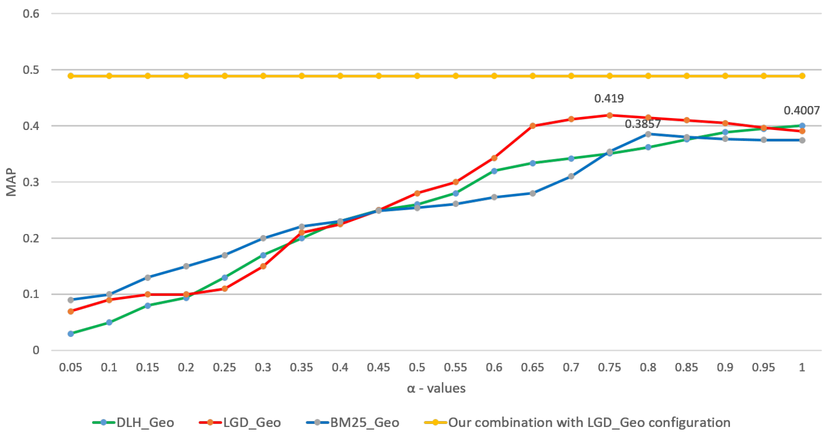

Figure 6.8: Evaluation of the proposed ranking combination strategy by comparing with the classical linear combination for different values of the $\alpha$ parameter and using GeoCLEF 2005

The Figure 6.9 shows the $MAP$ values obtained using all the combination ranking techniques previously described. It can be clearly appreciated that our combination gives the best result among all the others for any of the three configuration.

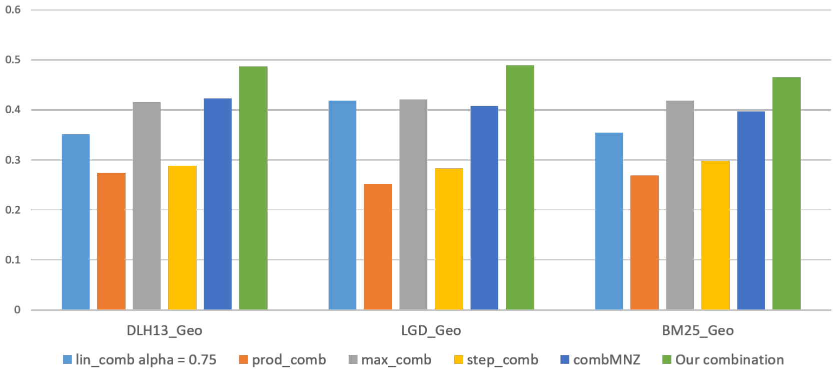

Figure 6.9: Comparison of all the described combination ranking strategies

## 6.5 Global Results Analysis

The goal of this last experiment is to evaluate the general performance of the proposed GIR system. The Tables 6.3 6.4 6.5 6.6 show a comparison of the *recall* and $MAP$ values among all the different configurations using for the evaluation the four GeoCLEF track campaigns. The first three rows of each table correspond to the results obtained by applying standard IR techniques without any geographic analysis, while the last three rows show the results using these standard techniques combined with the proposed geographic analysis. The tables clearly reflect that our approach improves all $MAP$ values demonstrating that adding the geographic analysis to traditional IR systems can improve the ranking results.

Table 6.3: Overall results using GeoCLEF 2005

| | **Model** | **Recall** | **MAP** |
|---|---|---|---|
| **Text** | *BM25* | 0.812 | 0.374 |
| | *DLH13* | 0.844 | 0.401 |
| | *LGD* | 0.856 | 0.391 |
| **Text+Geo** | *BM25_Geo* | 0.820 | **0.465** |
| | *DLH13_Geo* | 0.848 | **0.487** |
| | *LGD_Geo* | 0.862 | **0.489** |

Table 6.4: Overall results using GeoCLEF 2006

| | **Model** | **Recall** | **MAP** |
|---|---|---|---|
| **Text** | *BM25* | 0.787 | 0.263 |
| | *DLH13* | 0.756 | 0.235 |
| | *LGD* | 0.743 | 0.228 |
| **Text+Geo** | *BM25_Geo* | 0.772 | **0.323** |
| | *DLH13_Geo* | 0.756 | **0.291** |
| | *LGD_Geo* | 0.751 | **0.270** |

Table 6.5: Overall results using GeoCLEF 2007

| | **Model** | **Recall** | **MAP** |
|---|---|---|---|
| **Text** | $BM25$ | 0.844 | 0.266 |
| | $DLH13$ | 0.837 | 0.229 |
| | $LGD$ | 0.849 | 0.245 |
| **Text+Geo** | $BM25\_Geo$ | 0.852 | **0.319** |
| | $DLH13\_Geo$ | 0.840 | **0.266** |
| | $LGD\_Geo$ | 0.855 | **0.294** |

Table 6.6: Overall results using GeoCLEF 2008

| | **Model** | **Recall** | **MAP** |
|---|---|---|---|
| **Text** | $BM25$ | 0.733 | 0.281 |
| | $DLH13$ | 0.749 | 0.283 |
| | $LGD$ | 0.704 | 0.267 |
| **Text+Geo** | $BM25\_Geo$ | 0.736 | **0.345** |
| | $DLH13\_Geo$ | 0.753 | **0.329** |
| | $LGD\_Geo$ | 0.705 | **0.305** |

Another analysis was done based on the $p@5$ and $p@10$ evaluation metrics [9], which is defined as the *precision* obtained on the first 5 and 10 retrieved documents respectively. The Figure 6.10 shows how in all the cases the geographic analysis puts on the top ranked list more relevant documents which is desired result because in an scenario like the Web, users expect to find their information needs in the first ranking positions.

A comparison with other existing GIR system approaches was made also using GeoCLEF test collections. We remark that this analysis was made based on the results published on each GeoCLEF track. Tables 6.7,6.8,6.9,6.10 show the results of the comparison using GeoCLEF 2005 [41], 2006 [40], 2007 [69] and 2008 [68] track respectively. Each table includes the results obtained from the combinations $BM25\_Geo$, $DLH13\_Geo$ and $LGD\_Geo$ described above, and the three best results obtained by the teams participating in each GeoCLEF track.

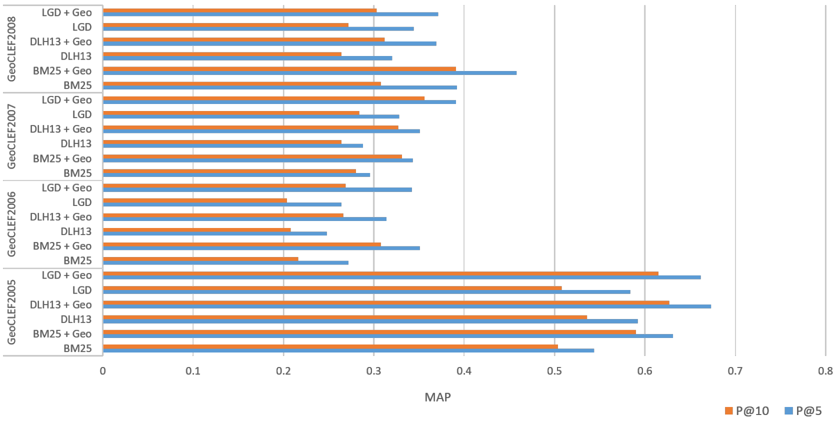

Figure 6.10: Comparison of all the described combination ranking strategies

Table 6.7: Comparison of $BM25\_Geo, DLH13\_Geo$, $LGD\_Geo$ with the three best results obtained in GeoCLEF 2005

| **Model** | **MAP** |
|---|---|
| $berkeley - 2\_BKGeoE1$ | 0.3936 |
| $csu - sanmarcos\_csusm1$ | 0.3613 |
| $alicante\_irua - en - ner$ | 0.3495 |
| $BM25\_Geo$ | **0.465** |
| $DLH13\_Geo$ | **0.487** |
| $LGD\_Geo$ | **0.489** |

Table 6.8: Comparison of $BM25\_Geo, DLH13\_Geo$, $LGD\_Geo$ with the three best results obtained in GeoCLEF 2006

| **Model** | **MAP** |
|---|---|
| $XLDBGeo$ | 0.3034 |
| $enTDpooled$ | 0.2723 |
| $SMGeoEN4notpooled$ | 0.2637 |
| $BM25\_Geo$ | **0.323** |
| $DLH13\_Geo$ | **0.291** |
| $LGD\_Geo$ | **0.270** |

Table 6.9: Comparison of $BM25\_Geo, DLH13\_Geo$, $LGD\_Geo$ with the three best results obtained in GeoCLEF 2007

| **Model** | **MAP** |
|---|---|
| $TALPGEOIRTD2$ | 0.285 |
| $BERKMOENBASE$ | 0.264 |
| $RFIAUPV06$ | 0.264 |
| $BM25\_Geo$ | **0.319** |
| $DLH13\_Geo$ | **0.266** |
| $LGD\_Geo$ | **0.294** |

Table 6.10: Comparison of $BM25\_Geo, DLH13\_Geo$, $LGD\_Geo$ with the three best results obtained in GeoCLEF 2008

| **Model** | **MAP** |
|---|---|
| $DFKIGEON3$ | 0.3037 |
| $SINAI_GC_MONO$ | 0.2664 |
| $XLDBTEAMEN6$ | 0.2755 |
| $BM25\_Geo$ | **0.345** |
| $DLH13\_Geo$ | **0.329** |
| $LGD\_Geo$ | **0.305** |

# Chapter 7

# Conclusions and Future Work

This thesis outlines an approach for a GIR solution. It combines standard IR techniques, for processing documents in a classical way, with a set of techniques oriented to the geographical analysis.

The geographical analysis is built on top of GeoNW, a new geographic knowledge base. This analysis includes effective toponym recognition and toponym disambiguation techniques.

We describe an extensive evaluation process of the main geographic analysis components as well as of the complete GIR solution. This evaluation process experimentally demonstrate the quality of this proposal.

We also remark the results of the proposed toponym recognition and disambiguation strategies, which are evaluated using TUD-Loc-2013 test collection. Results show that our algorithm outperforms recent Katz's strategies as well as the web services approaches in both aspects, TR and TD, obtaining higher values of precision, recall and F1 score. According to these results, the proposed toponym recognition and disambiguation algorithms can be used as part of further geographic information analysis applications, such as search engines.

The experiments also prove the effectiveness of GeoNW ontology. It was compared with GeoNames ontology using all four GeoCLEF campaigns always reflecting its suitability over GeoNames. This, in addition to the fact that the toponym recognition and disambiguation algorithms are built on top of GeoNW, makes the proposed geography ontology a valuable resource for processing geographic information.

The overall results also show that the whole proposal outperforms baseline techniques as well as other approaches reported by GeoCLEF campaigns. The best result is achieved by combining the *LGD* model with the proposed geographical analysis, obtaining a mean average precision of 0.489. These results suggest that the new strategy improves the accuracy of the retrieved document list according to the user's query and those also ratify the state of the art quality of the complete strategy.

The reliability of the results could be increased using more recent test collections. To the best of our knowledge, there has been no campaign to evaluate geographic information retrieval techniques since GeoCLEF 2008.

# Bibliography

[1] Rocío Abascal-Mena, Erick López-Ornelas, and J Sergio Zepeda. Geographic information retrieval and visualization of online unstructured documents. In *International Journal of Computer Information Systems and Industrial Management Applications*, volume 5, pages 89–97. MIR Labs, 2013.

[2] Alia I Abdelmoty, Philip D Smart, Christopher B Jones, Gaihua Fu, and David Finch. A critical evaluation of ontology languages for geographic information retrieval on the internet. *Journal of Visual Languages and Computing*, 16(4):331–358, 2005.

[3] Adeyinka K Akanbi. Lb2co: A semantic ontology framework for b2c ecommerce transaction on the internet. In *International Journal of Research in Computer Science*, volume 4, pages 1–9. White Globe Publications, 2014.

[4] Adeyinka K Akanbi, Olusanya Y Agunbiade, Sadiq Kuti, and Olumuyiwa J Dehinbo. A semantic enhanced model for effective spatial information retrieval. *arXiv preprint arXiv:1406.1969*, 2014.

[5] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280. ACM, 2004.

[6] Ivo Anastácio, Bruno Martins, and Pável Calado. A comparison of different approaches for assigning geographic scopes to documents. In *Proceedings of the 1st INForum-Simpósio de Informática*, pages 1–12. Citeseer, 2009.

[7] Leonardo Andrade and Mário J Silva. Relevance ranking for geographic ir. In *Workshop on Geographic Information Retrieval, SIGIR'06, Seattle, Washington, USA*, pages 1–4. ACM Press, 2006.

[8] M Andrea Rodriguez and Max J Egenhofer. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. In *International Journal of Geographical Information Science*, volume 18, pages 229–256. Taylor & Francis, 2004.

[9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval, 2nd Edition*, chapter Retrieval Evaluation, page 913. Addison-Wesley, 2011.

[10] Andrea Ballatore, David C. Wilson, and Michela Bertolotto. Computing the semantic similarity of geographic terms using volunteered lexical definitions. In *International Journal of Geographical Information Science*, volume 27. Taylor & Francis, 2013.

[11] Andrea Ballatore, David C Wilson, and Michela Bertolotto. A survey of volunteered open geo-knowledge bases in the semantic web. In *Quality issues in the management of web information*, pages 93–120. Springer, 2013.

[12] Kate Beard and Vyjayanti Sharma. Multidimensional ranking for data in digital spatial libraries. In *International Journal on Digital Libraries*, volume 1, pages 153–160. Springer, 1997.

[13] Imene Bensalem and Mohamed K Kholladi. Toponym disambiguation by arborescent relationships. *Journal of Computer Science*, 6(6):653, 2010.

[14] Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, and Yann Mathet. Geographic reference analysis for geographic document querying. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pages 55–62. Association for Computational Linguistics, 2003.

[15] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[16] Mike Blumenthal. Ed parsons: 1 in 3 searches at google are local, 2012.

[17] Conrad Bock, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, and Mike Smith. Owl 2 web ontology language structural specification and functional-style syntax. In *W3C Recommendation*, pages 1–63, 2012.

[18] Matthias Boehm, Benjamin Schlegel, Peter Benjamin Volk, Ulrike Fischer, Dirk Habich, and Wolfgang Lehner. Efficient in-memory indexing with generalized prefix trees. In *BTW*, volume 180, pages 227–246, 2011.

[19] Davide Buscaldi. Approaches to disambiguating toponyms. *SIGSPATIAL Special*, 3(2):16–19, 2011.

[20] Davide Buscaldi and Bernardo Magnini. Grounding toponyms in an italian local news corpus. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, pages 15:1–15:5. ACM, 2010.

[21] Davide Buscaldi and Paolo Rosso. Geo-wordnet: Automatic georeferencing of wordnet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA).

[22] Davide Buscaldi, Paolo Rosso, and Emilio Sanchis Arnal. Using the wordnet ontology in the geoclef geographical information retrieval task. In *GeoCLEF2005*, pages 939–946. Springer-Verlag, 2006.

[23] Davide Buscaldi and Paulo Rosso. A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science*, 22(3):301–313, 2008.

[24] Guoray Cai. Geovsm: An integrated retrieval model for geographic information. In *Geographic Information Science*, volume 2478 of *Lecture Notes in Computer Science*, pages 65–79. Springer, 2002.

[25] Guoray Cai. Visualization for geographical information retrieval. In *New Technology in Library and Information Services: Special issue on information visualization*, volume 10, 2004.

[26] Cláudio Elizio Calazans Campelo and Cláudio de Souza Baptista. A model for geographic knowledge extraction on web documents. In *Advances in Conceptual Modeling-Challenging Perspectives*, volume 5833 of *Lecture Notes in Computer Science*, pages 317–326. Springer, 2009.

[27] Nuno Cardoso and Mário J Silva. Query expansion through geographical feature types. In *Proceedings of the 4th ACM workshop on Geographical Information Retrieval, Lisbon, Portugal*, pages 55–60. ACM, 2007.

[28] Hsinchun Chen, Terence R. Smith, Mary Larsgaard, Linda L. Hill, and Marshall Ramsey. A geographic knowledge representation system for multimedia geospatial retrieval and analysis. In *International Journal on Digital Libraries*, volume 1, pages 132–152. Springer-Verlag, 1997.

[29] Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 277–288. ACM, 2006.

[30] Stéphane Clinchant and Éric Gaussier. Retrieval constraints and word frequency distributions a log-logistic model for ir. *Information retrieval*, 14(1):5–25, 2011.

[31] Paul Clough. Extracting metadata for spatially-aware information retrieval on the internet. In *Proceedings of the 2005 Workshop On Geographic Information Retrieval, Bremen, Germany*, pages 25–30. ACM, 2005.

[32] Gao Cong, Christian S Jensen, and Dingming Wu. Efficient retrieval of the top-k most relevant spatial web objects. In *Proceedings of the VLDB Endowment*, volume 2, pages 337–348. VLDB Endowment, 2009.

[33] Jean-Charles de Borda. Mémoire sur les élections au scrutin, histoire de l?académie royale des sciences. *Paris, France*, 1781.

[34] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th VLDB Conference, Cairo, Egypt*, pages 545–556. Morgan Kaufmann, 2000.

[35] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for*

*Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[36] Frederico T Fonseca, Max J Egenhofer, Peggy Agouris, and Gilberto Câmara. Using ontologies for integrated geographic information systems. In *Transactions in GIS*, volume 6, pages 231–257. Wiley Online Library, 2002.

[37] Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Building a geographical ontology for intelligent spatial search on the web. In *In Proceedings of IASTED International Conference on Databases and Applications*, pages 167–172. Spriner Verlag, 2005.

[38] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the First International Workshop on Location and the Web*, volume 300 of *ACM International Conference Proceeding Series*, pages 49–56. ACM, 2008.

[39] Eric Garbin and Inderjeet Mani. Disambiguating toponyms in news. In *In Proceedings of the Human Language Technology Conference (HLT-EMNLP'05*, pages 363–370. ACL, 2005.

[40] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, and Nicola Ferro. Geoclef 2006: The CLEF 2006 cross-language geographic information retrieval track overview. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers*, pages 852–876, 2006.

[41] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, and Paul D. Clough. Geoclef: the CLEF 2005 cross-language geographic information retrieval track. In *Working Notes for CLEF 2005 Workshop co-located with the 9th European Conference on Digital Libraries (ECDL 2005), Wien, Austria, September 21-22, 2005.*, 2005.

[42] Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. Geowordnet: A resource for geo-spatial applications. In *7th Extended Semantic Web Conference (ESWC)*, pages 121–136. Springer-Verlag, 2010.

[43] Michael F Goodchild. Geographical data modeling. In *Computers & Geosciences*, volume 18, pages 401–408. Elsevier, 1992.

[44] Ramaswamy Hariharan, Bijit Hore, Chen Li, and Sharad Mehrotra. Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems. In *Scientific and Statistical Database Management, 2007. SSBDM'07. 19th International Conference on*, pages 16–16. IEEE, 2007.

[45] Claudia Hauff, Dolf Trieschnigg, and Henning Rode. University of twente at geoclef 2006: geofiltered document retrieval. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 958–961. Springer, 2007.

[46] Gobe Hobona, Philip James, and David Fairbairn. Multidimensional visualisation of degrees of relevance of geographic data. *International Journal of Geographical Information Science*, 20(5):469–490, 2006.

[47] Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. The semantics of similarity in geographic information retrieval. In *Journal of Spatial Information Science*, pages 29–57, 2014.

[48] Christopher B Jones, Alia I Abdelmoty, David Finch, Gaihua Fu, and Subodh Vaid. The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science, Third International Conference, GIScience, Adelphi, MD, USA*, volume 3234 of *Lecture Notes in Computer Science*, pages 125–139. Springer, 2004.

[49] Christopher B Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. In *Spatial information theory*, pages 322–335. Springer, 2001.

[50] Christopher B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R.Weibel. Spatial information retrieval and geographical ontologies an overview of the spirit project. In *SIGIR Workshop on Geographic Information Retrieval.* ACM, 2002.

[51] Christopher B Jones and Ross S Purves. Geographical information retrieval. In *International Journal of Geographical Information Science*, volume 22, pages 219–228. Taylor & Francis, 2008.

[52] Philipp Katz and Alexander Schill. To learn or to rule: Two approaches for extracting geographic information from unstructured text. In *11th Australasian Data Mining Conference (AusDM'13)*, 2013.

[53] Carsten Kessler, Krzysztof Janowicz, and Mohamed Bishr. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 91–100. ACM, 2009.

[54] Donald E. Knuth. The art of computer programming. *Sorting and searching*, 3:426–458, 1999.

[55] Petr Kremen, Marek Smid, and Zdenek Kouba. Owldiff: A practical tool for comparison and merge of owl ontologies. In *22nd International Workshop on Database and Expert Systems Applications*, pages 229–233. IEEE Computer Society, 2011.

[56] R. R. Larson and P. L. Frontiera. Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, volume 3232 of *Lecture Notes in Computer Science*, pages 45–56. Springer-Verlag, 2004.

[57] Ray R. Larson. Geographic information retrieval and spatial browsing. In *Proceedings of the Data Processing Clinic - Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, pages

81–124. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1995.

[58] Robert Laurini. Geographic ontologies, gazetteers and multilingualism. In *Future Internet*, volume 7, pages 1–23. Multidisciplinary Digital Publishing Institute, 2015.

[59] Joon Ho Lee. Analyses of multiple evidence combination. In *ACM SIGIR Forum*, volume 31, pages 267–276. ACM, 1997.

[60] Jochen L Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, Institute of Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, 2007.

[61] Zhisheng Li, Ken CK Lee, Baihua Zheng, Wang-Chien Lee, Dik Lee, and Xufa Wang. Ir-tree: An efficient index for geographic document search. *IEEE Transactions on Knowledge and Data Engineering*, 23(4):585–599, 2011.

[62] Michael D Lieberman and Hanan Samet. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 731–740. ACM, 2012.

[63] Michael D Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 201–212. IEEE, 2010.

[64] Michael D Lieberman, Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. Steward: architecture of a spatio-textual search engine. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, page 25. ACM, 2007.

[65] Sha Lu, Ben He, and Jungang Xu. Hyper-geometric model for information retrieval revisited. In *Information Retrieval Technology*, volume 8281 of *Lecture Notes in Computer Science*, pages 62–73. Springer Berlin Heidelberg, 2013.

[66] Huanhuan Lv and Weidong Song. Semantic similarity measurement model based on geographic ontology. In *Journal of Information & Computational Science*, volume 10, pages 3869–3876. Binary Information Press, 2013.

[67] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. From puppy to maturity: Experiences in developing terrier. *Proc. of OSIR at SIGIR*, pages 60–63, 2012.

[68] Thomas Mandl, Paula Carvalho, GiorgioMaria Di Nunzio, Fredric Gey, RayR. Larson, Diana Santos, and Christa Womser-Hacker. Geoclef 2008: The clef 2008 cross-language geographic information retrieval track overview. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *Lecture Notes in Computer Science*, pages 808–821. Springer Berlin Heidelberg, 2009.

[69] Thomas Mandl, Fredric C. Gey, Giorgio Maria Di Nunzio, Nicola Ferro, Ray R. Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. Geoclef 2007: The CLEF 2007 cross-language geographic information retrieval track overview. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 745–772. Springer, 2007.

[70] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[71] Bruno Martins and Pável Calado. Learning to rank for geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, page 21. ACM, 2010.

[72] Bruno Martins and Mário J Silva. A graph-ranking algorithm for geo-referencing documents. In *2013 IEEE 13th International Conference on Data Mining*, pages 741–744. IEEE Computer Society, 2005.

[73] Bruno Martins, Mário J Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34. ACM, 2005.

[74] Bruno Martins, Mário J Silva, and Leonardo Andrade. Indexing and ranking in geo-ir systems. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34. ACM, 2005.

[75] Christian Middleton and Ricardo Baeza-Yates. A comparison of open source search engines, 2007.

[76] George A Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[77] Daniel R Montello. The geometry of environmental knowledge. In *Theories and methods of spatio-temporal reasoning in geographic space*, pages 136–152. Springer, 1992.

[78] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[79] NIST. Common evaluation measures. In *Proceedings of the 15th Text Retrieval Conference (TREC 2006)*, 2006. Appendix to the 15th TREC proceedings.

[80] Simon Overell, João Magalhães, and Stefan" Rüger. *Forostar: A System for GIR*, pages 930–937. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[81] Simon E Overell. *Geographic information retrieval: Classification, disambiguation and modelling.* PhD thesis, Imperial College London, University of London, 2009.

[82] Damien Palacio, Curdin Derungs, and Ross Purves. Development and evaluation of a geographic information retrieval system using fine grained toponyms. *J. Spatial Information Science*, 11(1):1–29, 2015.

[83] Damien Palacio, Christian Sallaberry, and Mauro Gaio. Normalizing spatial information to better combine criteria in geographical information retrieval. In *ECIR, 31st European Conference on Information Retrieval*, pages 37–49, 2009.

[84] Jeong-Hoon Park. Spatial semantic search in location-based web services. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 9–14. International World Wide Web Conferences Steering Committee, ACM, 2014.

[85] Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, et al. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 53–58. European Language Resources Association (ELRA), 2006.

[86] Ross S Purves, Paul Clough, Christopher B Jones, Avi Arampatzis, Benedicte Bucher, David Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, et al. The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. *International journal of geographical information science*, 21(7):717–745, 2007.

[87] Kirk Roberts, Cosmin Adrian Bejan, and Sanda M Harabagiu. Toponym disambiguation using events. In *FLAIRS Conference*, volume 10, page 1, 2010.

[88] João B Rocha-Junior, Orestis Gkorgkas, Simon Jonassen, and Kjetil Nørvåg. Efficient processing of top-k spatial keyword queries. In *International Symposium on Spatial and Temporal Databases*, pages 205–222. Springer, 2011.

[89] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.

[90] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[91] M Sanderson and J Kohler. Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval*, volume 2, pages 1–5. Sheffield, 2004.

[92] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 105–108, 1994.

[93] Mário J Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding geographic scopes to web resources. In *Computers, Environment and Urban Systems*, volume 30, pages 378–399. Elsevier, 2006.

[94] Adam Rae Simon Overell and and Stefan Rger. Geographic and textual data fusion in forostar. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008*, volume 5706, pages 838–842. Springer Berlin Heidelberg, 2008.

[95] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 127–136, London, UK, UK, 2001. Springer-Verlag.

[96] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1297–1304, 2004.

[97] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006.

[98] Nicola Stokes, Yi Li, Alistair Moffat, and Jiawen Rong. An empirical study of the effects of nlp components on geographic ir performance. In *International Journal of Geographical Information Science*, volume 22, pages 247–264. Taylor & Francis, 2008.

[99] Tomek Strzalkowski. Natural language information retrieval. In *Information Processing and Management*, volume 31, pages 397–417. Elseiver, 1995.

[100] Jiayu Tang and Mark Sanderson. Evaluation and user preference study on spatial diversity. In *European Conference on Information Retrieval*, pages 179–190. Springer, 2010.

[101] Benjamin E Teitler, Michael D Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: a new view on news. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM, 2008.

[102] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[103] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. In *The knowledge engineering review*, volume 11, pages 93–136. Cambridge Univ Press, 1996.

[104] Marc van Kreveld, Iris Reinbacher, Avi Arampatzis, and Roelof van Zwol. Distributed ranking methods for geographic information retrieval. In *Developments in Spatial Data Handling*, pages 231–243. Springer, 2005.

[105] U. Visser, H. Stuckenschmidt, G. Schuster, and T. Vogele. Ontologies for geographic information processing. In *Computers & Geoscience*, volume 28, pages 103–117. Elseiver Science, 2002.

[106] Ubbo Visser, Thomas Vögele, and Christoph Schlieder. Spatio-terminological information retrieval using the buster system. *Proc. of Environmental Informatcs*, 2002.

[107] Wei Wang, Jiong Yang, and Richard Muntz. Pk-tree: a spatial index structure for high dimensional point data. In *Information organization and databases*, pages 281–293. Springer, 2000.

[108] Marc Wick. Geonames. `http://www.geonames.org/`, 2005. [Last accessed Dec-2016].

[109] Marc Wick. Geonames google group. `https://groups.google.com/forum/#!topic/geonames/RKcANM4ZvyA`, 2007. [Last accessed Dec-2016].

[110] Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964. Association for Computational Linguistics, 2011.

[111] Allison Gyle Woodruff and Christian Plaunt. Gipsy: Georeferenced information processing system. In *Journal of the American Society for Information Science*, volume 45, pages 645–655. Citeseer, 1994.

[112] Bo Yu and Guoray Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 49–54. ACM, 2007.

[113] Bo Yu and Guoray Cai. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 49–54. ACM, 2007.

[114] Yinghua Zhou, Xing Xie, Chuang Wang, Yuchang Gong, and Wei-Ying Ma. Hybrid index structures for location-based web search. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 155–162, New York, NY, USA, 2005. ACM.

[115] Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '05, pages 354–362, New York, NY, USA, 2005. ACM.