Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Sidra5: A Search System with Geographic Signatures

David José Vaz Cruz

Mestrado em Engenharia Informática

2007

Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Sidra5: Scalable Search System with Geographic Signatures

David José Vaz Cruz

Projecto orientado pelo Prof. Dr. Mário J. Gaspar da Silva

Mestrado em Engenharia Informática

2007

Resumo

Este trabalho consistiu no desenvolvimento de um sistema de pesquisa de informação com raciocínio geográfico, servindo de base para uma nova abordagem para modelação da informação geográfica contida nos documentos, as assinaturas geográficas. Pretendeu-se determinar se a semântica geográfica presente nos documentos, capturada através das assinaturas geográficas, contribui para uma melhoria dos resultados obtidos para pesquisas de cariz geográfico. São propostas e experimentadas diversas estratégias para o cálculo da semelhança entre as assinaturas geográficas de interrogações e documentos. A partir dos resultados observados conclui-se que, em algumas circunstâncias, as assinaturas geográficas contribuem para melhorar a qualidade das pesquisas geográficas.

PALAVRAS-CHAVE:
Pesquisas geográficas, indexação, ordenação, assinaturas geográficas

Abstract

The dissertation report presents the development of a geographic information search system which implements geographic signatures, a novel approach for the modeling of the geographic information present in documents. The goal of the project was to determine if the information with geographic semantics present in documents, captured as geographic signatures, contributes to the improvement of search results. Several strategies for computing the similarity between the geographic signatures in queries and documents are proposed and experimented. The obtained results show that, in some circunstances, geographic signatures can indeed improve the search quality of geographic queries.

KEYWORDS:
Geographic searches, ranking, indexing, geographic signatures

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The information search paradigm has become an essential part of our lives. Search systems are everywhere, from search engines, e-mail applications and web web browsers, to desktop search applications. We are overflowen by information and the world wide web is one of the main drivers. The web has experienced exponential growth and shows no sign of slowing down. This simple observation reveals the growing dependence of search engines on sustaining our growing need for information. We need tools to ease our information mining needs.

Search engines must constantly evolve to be able to accommodate an increasingly larger amount of information published on the web, while responding at the same time to the users' growing demand for more precise and accurate search results.

In order to return relevant results, matching users' expectations, it becomes necessary to perceive the information need using a few input words. It is crucial that the context in which the query appears is understood, so that search efforts can focus in the information that conforms to that context. In this effort of contextualization, a query analysis performed by Sanderson and Kohler [2004] showed that around 15% of search engine queries have a well defined geographic context, and Zhang et al. [2006] showed that geographic queries have in average a higher number of characters and terms. The presence of more information in geographic queries, which in part has a more restrictive context, enables potentially improved results.

Geographic queries open up a new set of opportunities, not without an all new set of difficulties and challenges. They require the creation of representations of geographic knowledge for machine reasoning; understanding the geographic scopes of web documents and disambiguating among possible conflicting meanings for a geographic term; creating new indexing schemas that account for the inclusion of geographic information; developng new ranking algorithms that measure the geographic relevancy of the documents relatively to the query; extracting the geographic information, explicitly or implicitly, submitted by the users in queries.

Geographic Information Retrieval(GIR) emerged as the study of the retrieval and processing of geographic information. As an outcome of the research in GIR of the XLDB group, the Tumba! search engine, initially a purely textual search engine, was extended with geographic capabilities. Impelled by the participations on GeoCLEF [Martins et al., 2006][Gey et al., 2006], the GeoCLEF track of CLEF, this year Tumba! applied a new approach to model, index, rank and retrieve geographic information, representing all the geographic information present on queries and documents in data structures designated as *geographic signatures*. Geographic Signatures represent all the geographic references, and respective confidence measures, contained in a document or in a query.

In the research environment where Tumba! is being developed, new algorithms are constantly surfacing, new techniques must be tested and new usage scenarios must be considered. This constant need to catch-up with innovations on topics related to the core research of the group requires a substantial amount of effort and limits the amount of time that can be spared on system maintenance tasks.

The recency of the GIR research area allows a whole spectrum of opportunities. the use of geographic information to surpass textual retrieval techniques is largely untackled, making GIR an atractive research subject.

The scope of the geographic signatures approach is not confined to academic research. As it matures, it will become a concrete functionality of the Tumba! search engine, which will be used to support geographic searches submited by users.

This work describes the use of the geographic information as geographic signatures by *Sidra*, the component that is responsible for the indexing and ranking processes of the Tumba! search engine. It also describes the integration of the Sidra in the Tumba! search engine and in the XLDB GeoCLEF framework, a prototype built for evaluation purposes that was used by the XLDB Group in its 2007 participation of GeoCLEF.

1.1 Objectives

For the development of the new Sidra version, version 5, the following objectives were identified:

Geographic Relevance based on Geographic Signatures: the new version of *Sidra* should implement a novel approach for searching geographic information, based on *geographic signatures*.

This approach intends to overcome the limitations detected in previous approaches, which lead to the observation that the premise of one geographic scope per document was too vulnerable to the wrong assignment of geographic

features to documents (geographic features being conceptualizations of physical and administrative locations), and was sometimes too restrictive[Martins et al., 2006].

Geographic Signatures Search versus Textual Search: determine if the proposed approach of representing geographic information and calculating the geographic similarity of documents using geographic signatures could obtain better search results than purely textual searches.

Geographic Scoring Strategies Comparison: implement and evaluate geographic scoring strategies based on geographic signatures for their reasoning.

Reduce Maintenance Efforts: the new version of *Sidra* should bring a reduction of required maintenance efforts, both at the administrative level and at the software development level, simplifying the configuration and installation of the Tumba! search engine, by favoring the use of existing and well documented modules to perform specific tasks.

Sidra5 was built in a more modular fashion, allowing a lower effort on the implementation of new relevance algorithms. Taking advantage of the rewriting of this component, a new textual indexing and ranking software module was selected to be included in Sidra5.

1.2 Document Structure

The rest of this thesis has the following structure:

Chapter 2 – reviews some of the main works in *Geographic Information Retrieval*, divided into four categories: information extraction and disambiguation, data structures and indexing, ranking and search, user interface.

Chapter 3 – rediscovers the research path that led to the formulation of the geographic signatures approach. It gives examples of deficiencies of previous approaches that geographic signatures aim to surpass. This chapter ends with a description of the evaluation measures for search quality, which are used in this work.

Chapter 4 – presents the architecture of Sidra5, describing its components and detailing the integration of Sidra into the Tumba! search engine and into the test prototype used for evaluations on GeoCLEF 2007. It also provides a comparative analysis of a set of existing *Information Retrieval* libraries, with the purpose of selecting one of them to be included in the next version of Sidra

as the textual indexing and ranking component. Further tests on the indexing and search quality of the libraries are also described.

Chapter 5 – details the architecture for a geographic search system based on geographic signatures, describing the structure of geographic signatures and the processes of geographic indexing, relevance calculation and search.

Chapter 6 – evaluates the geographic signatures approach using the topics of the 2007 edition of GeoCLEF, comparing several geographic scoring strategies in multiple query expansion scenarios. This chapter also details the results obtained using geographic signatures with the results of a purely textual approach.

Chapter 7 – recapitulates the initially proposed objectives and discusses the extent of their achievements and outilnes the conclusions on the use of geographic signatures on geographic search systems. The last chapter also ends with some directives for future works dedicated to extending and improving the geographic signatures approach.

Chapter 2

Related Works

As a sub-area of *Information Retrieval* (IR), *Geographic Information Retrieval* (GIR) has lately received a considerable attention by the scientific community.

There are two distinct approaches to the challenge of geographic information search, intended to respond to distinct necessities: GIR and *Geographic Information Systems* (GIS).

There is sometimes a lack of understanding of what differentiates these two disciplines, originating doubts about the necessity and utility of the existence of GIR processes. Although both are devoted to the retrieval of geographic information, their objectives and methods differ. Essentially, the difference between GIS and GIR is similar to the difference between relational databases and classic IR systems. GIS are devoted to the capture, management, analysis and presentation of geographically-related information, resorting to data structures that enable choosing the entities that match some logical selection criteria. The goal of GIR is, given a geographic context, to recover the information considered as potentially relevant, by order of relevance, in the most efficient possible way.

In search engine contexts, the concern is on the efficiency in the retrieval of relevant results and the quality of those results. As such, GIR has been driven by the search engine necessities to augmentate the quality of their services to users, trying to contextualize and adjust to their momentary needs.

A diversity of works about GIR have appeared recently. These works can be divided into four areas:

- Extraction and disambiguation of information
- Data structures and indexing
- Ranking and querying
- User interfaces

2.1 Extraction and Desambiguation of Information

Extraction refers to the tasks of identifying geographic information contained in web documents and/or information contextually relevant to the geographicity of the document. Desambiguation refers to the resolution of ambiguous geographic information, such as identical names for distinct localities.

The beginning of any geographic retrieval task resides in identifying the geographic information present in web documents (*extraction*).

Silva et al. [2006] presented a methodology for the attribution of geographic contexts to documents of the Portuguese web using an ontology that maps the geographic knowledge as geographic concepts and their relationships.

Yi Li and Cavedon [2006] and Markowetz et al. [2005] assigned multiples geographic contexts to web documents. This approach, based on probabilities, enables the association of documents to mutiple geographic contexts.

Julien Lesbegueries and Gaio [2006] assigned geographic contexts to textual units (such as sections, paragraphs and sentences) and tryed to infer geographic patterns inside documents with this information.

Qi Zhang and Ma [2006], aside the association of geographic contexts to documents, also calculated *serving scopes*. The *serving scope* corresponds to the serving area that a given resource has in a given geographic context. For example, the *serving scope* of a query for "Restaurants in Lisbon" will be substantially smaller that a query for "Airports in Lisbon".

In Graupmann and Schenkel [2006], all the geographic references that exist in the documents are preserved. When a query is performed on the documents, only the geographic references that arise in a context similar to the query context are used for the purpose of ranking calculation.

Tezuka et al. [2006] presented a system that identifies the geographic contexts of documents and associates additional external information that refers to the same geographic context.

Souza et al. [2005] discussed the contribution of gazetteers to the identification and assignment of geographic meaning to documents, focusing particularly on the benefit of ontologically enhanced gazetteers in the discovery of the geographic context present on documents.

Walker et al. [2005] proposed and evaluated spatial Bayesian network algorithms as a mean for automatically infering the relationship between geographic contexts.

2.2 Data Structure and Indexing

The use of geographic information spawned the development of new data structures that profit from this additional information in the indexing phase.

Zhou et al. [2005] describe comparative tests to three distinct hybrid spatial indexing strategies which combine textual and geographic information, concerning to the performance and information storage cost.

Zhisheng Li [2006] used two types of indexes. One index maps the hierarchical relation between explicit and implicit locations allowing the inference of locations relevant to a given search context. The other index associates the locations to a grid, enabling search by proximity.

Andrade and Silva [2006a] presented geographic information indexing structures. These structures separate the geographic and textual information, allowing the search of only one type of information.

Martins et al. [2005] presented index structures that benefit from several data structures: inverted indexes, spatial indexes and geographic ontologies.

2.3 Ranking and Querying

The availability of additional geographic information can be used by the IR ranking algorithms. To profit from this additional information, new ranking algorithms have been created. Also, it is now essential to identify which queries are geographic.

Andrade and Silva [2006b] presented a ranking schema to relate two locations. They also evaluated several strategies for combining textual and geographic information and the benefits of these approaches.

Markowetz et al. [2005] presented a ranking algorithm which intersects textually relevant documents with the documents' *geographic footprints*, to determine which are relevant to a query.

Zhang et al. [2006] analysed the structure of user submitted queries and how users rewrite queries to adjust the geographic context of submitted queries.

Bruno Martins and Afonso [2006] described a methodology to interpret and disambiguate user submitted queries using a geographic ontology.

Chen et al. [2006] analyzed the scalability of several algorithms for geographic queries processing, evaluating the efficiency of several strategies to combine textual and geographic information.

2.4 User Interfaces

Since the final objective of a geographic search system is to be useful to its users, the user interface is crucial to its acceptance and success. In this domain, new ways

to present geographic information have been researched, as well as novel forms of user interaction for the input of geographic information in queries.

Ware et al. [2002] presented a solution to resolve the incoherences that appear on maps as its scale decrease.

Tezuka et al. [2006] presented a proactive geographic search system, heavily inspired on car radios. The user chooses a query type (e.g., restaurants) and the results are constantly being updated as the position of the vehicle, received through GPS, changes.

Carmo et al. [2005] describe a prototype of a geographic referenced information system which allows the filtering of the information presented to the users according to some semantic criteria.

Hobona et al. [2005] presented a 3D interface for the presentation of results. The results entry are placed along three axis on semantic, temporal and geographic relevance.

2.5 Conclusion

GIR is a broad research area with very diversified works. It is rapidly evolving due to the ever growing internet demand for contextually more precise information. Yet, GIR still has to prove its efficacity relatively to the classic IR.

Markowetz et al. [2005] gave a complete view of how to integrate techniques of information extraction and disambiguation, indexing and search to create a completely functional geographic search engine. The approach used in the prototype for the modeling of the documents' geographic information, called *geographic footprints*, considers that a document can contain (and they usually do) more than one geographic reference and that all of them have to be considered. This idea is similar to the one used for the geographic signature approach described in this work, but its advantages still have to be demonstrated.

Chapter 3

From Textual Retrieval to Geographic Signatures

Textual Retrieval has been the standard for search technologies. Despite its wide spread utilization, the use of words as search tokens limit its overall effectiveness. Neither the geographic semantics nor the distance and hierarchical notions are interpreted and considered. This provides a window of opportunity for GIR techniques to surpass their textual homologous.

This chapter outlines the past research on GIR by the XLDB group, from the textual retrieval to geographic signatures. Section 3.1, summarily describes the used approaches and explains the motivation for each evoluationary iteration. Section 3.2, identifies the evaluation metrics used, during all this work, to compare the search quality of different algorithms and approaches.

3.1 Approaches

The retrieval approaches that have been the object of past research on the XLDB group for processing geographic queries are:

Text Retrieval: this approach has no explicit geographic processing. It partitions the document into tokens, usually words, and uses probabilistic algorithms to match the tokens of the documents with those present in the queries.

One geographic feature by document: this approach gathers all the geographic references contained in a document and assigns, as the geographic scope of the document, the geographic feature that best describes all the contained references[Martins et al., 2006].

Geographic Signature: this approach gathers all the distinct geographic references contained in the document into a data structure (geographic signature).

	Document
d_1	"When visiting **Portugal**, the restaurants that you cannot miss are ..."
d_2	"...eated, last night, in a small restaurant of **Bragança** called ..."
d_3	"You can visit our other restaurants in **Oporto**, **Bragança**, ..."
d_4	"...where you can taste the finest seafood that **Lisbon** has to offer. This restaurant ..."

(a) documents

	Queries
q_1	Restaurants in **Portugal**
q_2	Restaurants in **Bragança**
q_3	Restaurants in **Bragança** and **Oporto**

(b) queries

Table 3.1: Examples of document excerpts and queries

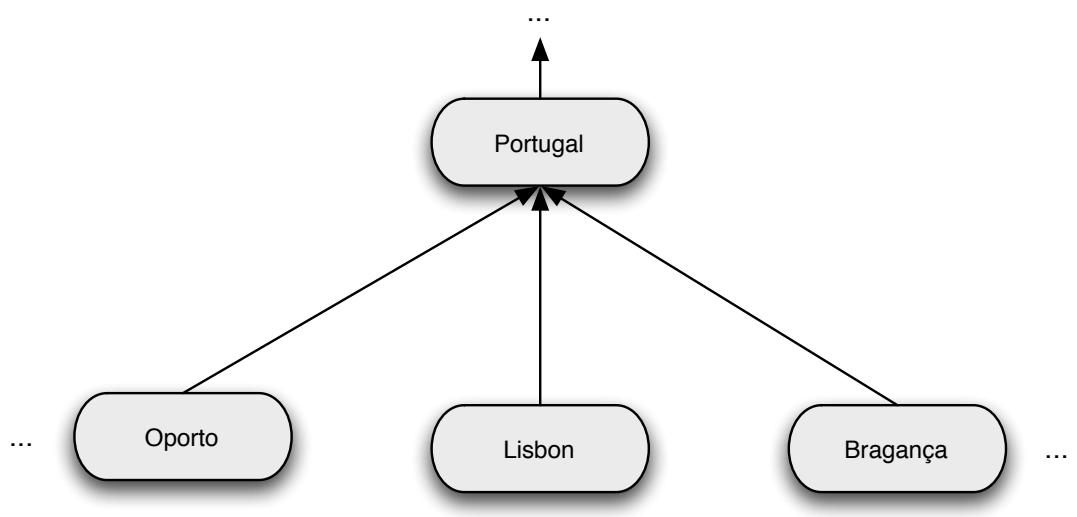


Figure 3.1: Example of geographic hierarchy subset

The geographic scope of the document is then composed of all of these references, each with a weight value that describes their importance in the geographic signature.

To ease the understanding of the motives that led to the successive evolutions, some fictional document excerpts (see Table 3.1a) and queries (see Table 3.1b) were used as examples. Additionally, Figure 3.1 contains a graph which represents the hierarchy of the geographic concepts contained in both the query and document examples.

Approach	q_1	q_2	q_3
Text Retrieval	d_1	d_2, d_3	d_2, d_3
One Geo Feature	d_1, d_2, d_3, d_4	d_2	d_1, d_2, d_3, d_4
Geo Signature	d_1, d_2, d_3, d_4	d_2, d_3	d_2, d_3

Table 3.2: Search quality test specification

Since textual retrieval has no notion of what is "geographic information" (such as geographic features, geographic names, spatial and hierarchical relations), it is only able to retrieve documents that explicitly contain the query terms. For this reason, in q_1, textual retrieval only recovers d_1 despite d_2, d_3 and d_4 being potentially relevant, since they are all siblings of the node referred in the query (Portugal).

To overcome the lack of hierarchical notion, the "one geographic feature by document" approach was created. However, this may be too simplistic, since "one geographic feature by document" adjusts the document scope to the geographic feature which best encompass all the geographic references contained in the document. For example, in q_2, "one geographic feature by document" shows a lower retrieval quality when compared to the textual search (less documents retrieved). In q_3, too many documents are retrieved since d_4 is mistakenly retrieved. This is caused by over-generalization of the query.

The geographic signature approach came as a natural evolution. It retrains the hierarchical notion that allows to retrieve all the correct documents for q_1, yet its finner granularity allows to answer correctly to q_3.

Theorically, the geographic signatures approach benefits from the notion of geographic knowledge, without suffering the shortcomings of "one geographic feature by document".

3.2 How to Measure?

When faced with a multitude of approaches and strategies to rank information according to multiple search criteria, it is necessary to evaluate then equally. In IR, there are several distinct measures to quantify the performance of search systems. This work will focus on three of them: precision, recall and MAP (Mean Average Precision).

The precision (P_r) is the fraction of the top r ranked documents that are relevant to the query. Precision is given by the following formula:

$$P_r = \frac{|\{relevant\ documents\}| \cap |\{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{3.1}$$

The recall (R_r) is the proportion of the total number of relevant documents that

were retrieved in the top r. Recall is given by the following formula:

$$R_r = \frac{|\{relevant\ documents\}| \cap |\{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{3.2}$$

These two measures can be combined into a measure called *interpolated precision*. It corresponds to the maximum precision value at (typically) 11 equidistant cut points from 0 percent to 100 percent of recall. Since recall is a non-decreasing function of rank the values of the interpolated precision are non-increasing as recall increases.

The MAP measure is the mean value of the average precisions computed for each of the queries separately. Average precision emphasizes returning more relevant documents earlier. It is the average of precisions computed after truncating the results list after each of the relevant documents in turn:

$$AveragePrecision = \frac{\sum_{r=1}^{N}(P_r \times rel(r))}{|\{relevant\ documents\}|} \tag{3.3}$$

where r is the rank, N the number of retrieved results, $rel()$ a binary function on the relevance of a given rank.

In today's competitions, such as CLEF, the MAP value is the main evaluation measure to evaluate the performance of a ranking algorithm.

Conceptually, the geographic signature approach has the potential to surpass the textual retrieval approach. Geographic signatures have to be tested to see if this assumption still holds in real world scenarios. In Chapter 5, the geographic signature specifics are presented and, in Chapter 6, this approach is evaluated.

The next Chapter describes the first software module which implements this novel approach, Sidra5, and describes its integration into two systems.

Chapter 4

Sidra5

Sidra has been used as a component of Tumba!, the search engine created by the XLDB research group, and in the evaluation system used by XLDB in the 2007 Geographic Track of CLEF (Cross Language Evaluation Forum), also known as GeoCLEF[?].

Both in Tumba! and on the GeoCLEF prototype, Sidra assumes the role of indexing and ranking component, even if these two systems have different purposes and architectures. Sidra creates the textual and geographic indexes and does the geographic reasoning, comparing the similarity between queries and documents.

This chapter lists the requirements of Sidra (Section 4.1), details its architecture (Section 4.2), describes the indexes structure (Section 4.3), presents how it was embedded into the two systems (Section 4.4 and 4.5) and describes the selection of the text IR framework used in Sidra (Section 4.6).

4.1 Requirements

The following requirements for the new version of Sidra were defined:

Implement Geographic Signatures: The main requirement for the new version of Sidra was to support the novel geographic signature approach, described in this dissertation.

Integration: Since both the systems where Sidra5 is integrated heavily use components written in Java, design decision had to be made to ensure a correct integration, favoring software modules written in Java. Additionally, Sidra had to seamlessly run in Linux operative system since the systems where it was integrated require this platform.

Software Reuse: Specific tasks of Sidra are implemented using existing software to achieve bigger modularity, software quality, support and fasten the development process. The use of open-source software was favored.

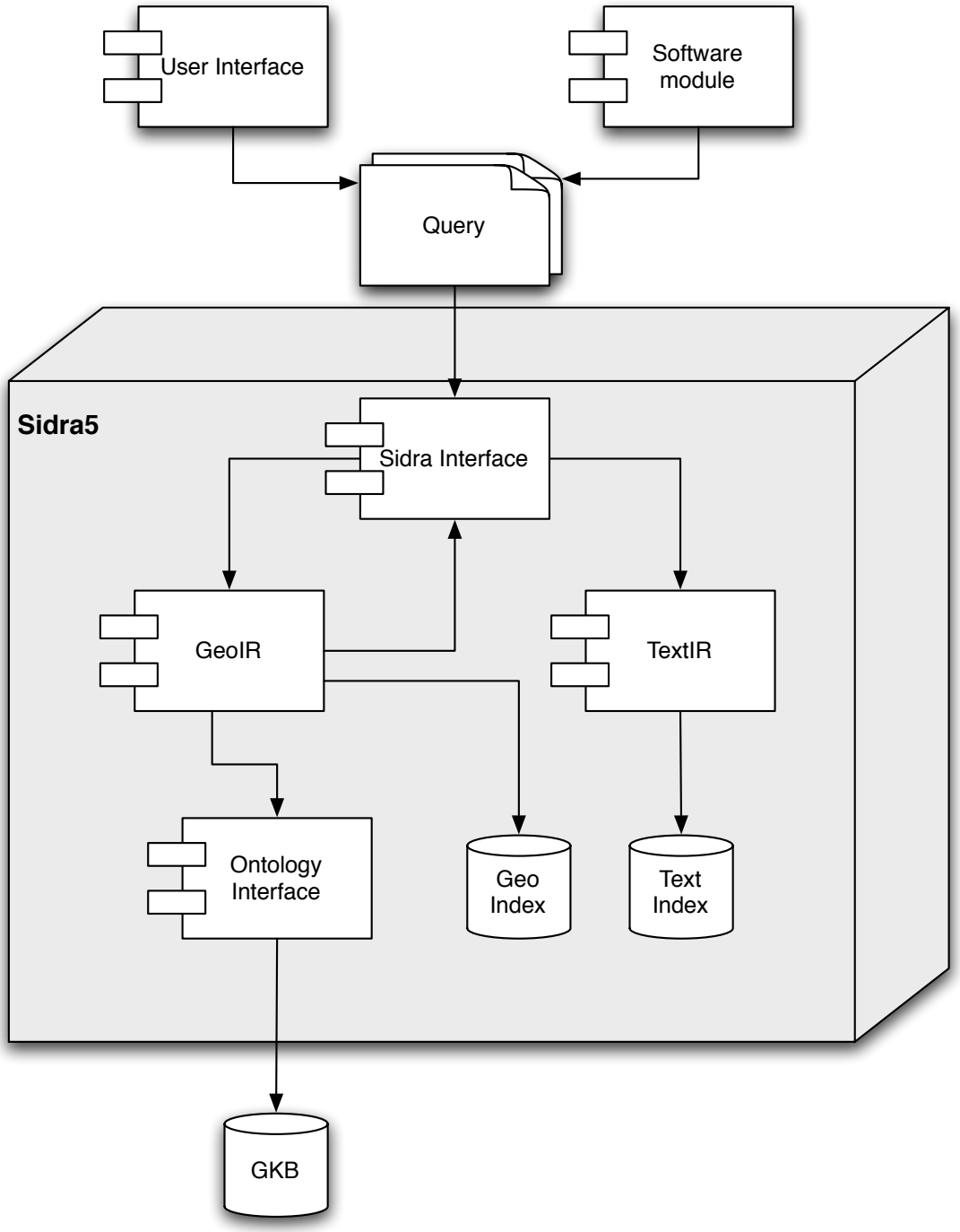


Figure 4.1: Sidra5 architecture

Geographic Search "on demand": Even if 15% of user submitted searches are geographic queries[Sanderson and Kohler, 2004], which justifies the current effort around GIR, the fact is that the majority of queries are not geographic. The new version of *Sidra* had to ensure that only geographic queries had to

support the additional computational effort required for the geographic reasoning and ensure that the appropriate calculation of relevance was performed, given the type of input query.

4.2 Architecture

Sidra is a software component with well defined purpose and functions, where the geographic indexing and ranking processes are decomposed into distinct tasks. In the newly developed version of Sidra, version 5, the architecture has been designed to emphasize flexibility and modularity, assigning indexing or ranking tasks to interchangeable sub-components. The internal architecture of Sidra5, shown in Figure 4.1, has the following units:

Sidra Interface: This module exposes the functionalities of Sidra5 to external components. The interface was created with modularity in mind, providing indexing and ranking functions. In this module incoming queries are routed to geographic retrieval or textual retrieval evaluation algorithms.

TextIR: This module offers text retrieval methods. MG4J[1], a software library that provides high indexing and querying efficiency and scalability, while supporting state of the art IR ranking algorithms, was chosen as basis for implementing the text IR framework of Sidra.

GeoIR: This sub-component offers the geographic reasoning methods. It is responsible for the geographic indexing and retrieval of signatures.

This module is implemented on top of the *Berkeley DB*–java edition[2].

It is also in this module that the geographic similarities between queries and documents are computed.

Ontology Interface: This component provides an interface with the Geographic Knowledge Base (GKB) [Chaves et al., 2007]. It offers methods to retrieve the information about the geographic entities in the ontology, which is used by heuristics for the computation of geographic similarities.

4.3 Indexes Structure

As depicted in Figure 4.1, Sidra5 uses two kinds of indexes: a textual index and a geographic index. The textual index consists of an inverted index `[term→<docid>]`,

[1] http://mg4j.dsi.unimi.it/
[2] http://www.oracle.com/database/berkeley-db/je/index.html

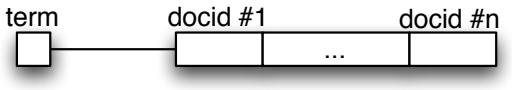


Figure 4.2: Text Index Structure

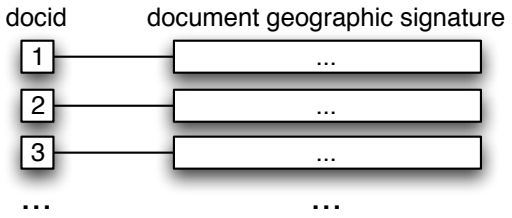


Figure 4.3: Geographic Index Structure

where *docid* is an unique document identifier. The geographic index is a forward index `[docid→geographic signature]` (see Figure 4.2 and 4.3 for representations of the text and geographic indexes, respectively). The motivation for this index division is that it allows simultaneous handling of both textual and geographic query types, i.e. queries with and without geographic context, in the most efficient way. If a query is not geographic, only the textual index is used, but if a geographic query is submitted both indexes are used.

For this division to succeed, consistency of the document identifiers in both indexing structures must be ensured. Each document has to have its own identifier and that identifier has to be unique, making it possible to connect the information of the textual index with the information of the geographic index.

With the uniqueness of the documents identification it is then possible to construct the two indexes in a parallel and independent fashion. With this index structure it is also possible to reconstruct just one of the two indexes without interfering with the normal behavior of the other one.

Geographic indexing in Sidra is performed in two phases, which can be run in parallel since the uniqueness of document identifiers (docid) is preserved.

1. The text indexing sub-module of *Sidra* is fed with the documents of the collection, creating the text index.

2. The geographic indexing sub-module of *Sidra* is fed with the geographic signatures of documents. It then creates the geographic index.

4.4 Sidra5 in the Tumba! Search Engine

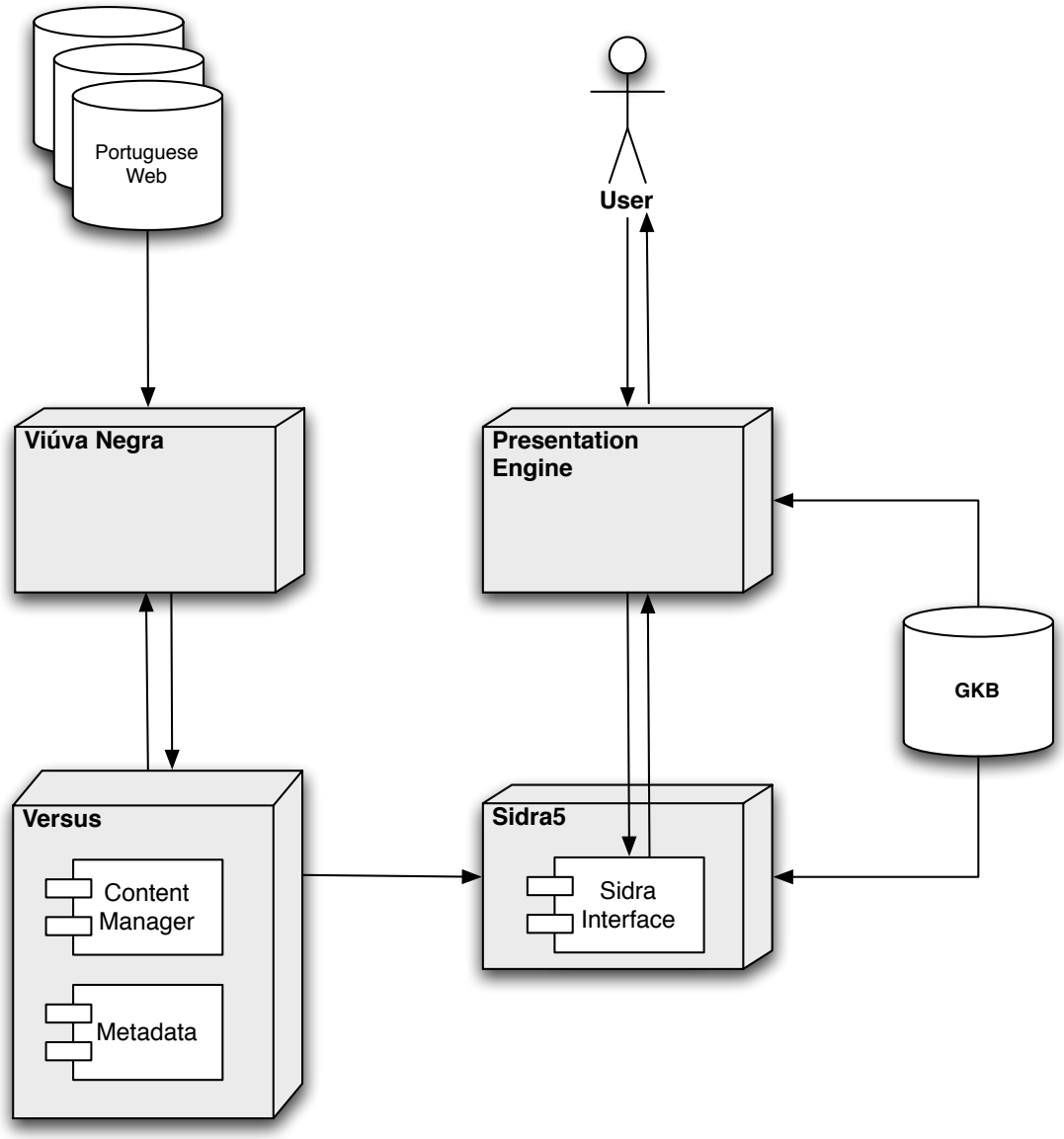


Figure 4.4: Tumba! Search Engine Architecture

The purpose of Tumba! is to provide scalable search capabilities across documents of the Portuguese web. In addition to Sidra, Tumba! has the following main components (see Figure 4.4):

Viúva Negra is a web crawler developed internally by the XLDB Group [Gomes, 2001, Gomes and Silva, 2001]. Viúva Negra aims to be a highly efficient, parallelizable, web crawler that has load balancing.

Versus is a repository for webcontent that provides high performance structured access to meta-data and extensible storage space for contents [Gomes et al., 2004]. It is composed by the *Content Manager* and the *Catalog*. The *Content Manager* provides storage space for the contents and mechanisms for the elimination of duplicates. The *Catalog* provides high performance access to structured meta-data. It keeps information about each content such as the date when it was collected and the reference to the location where it was stored in the *Content Manager*.

Presentation Engine is the component in the Tumba! search engine that does an initial analysis of the user's submitted queries, spell-checking the query, passing it to Sidra, and finally generating the web pages and presenting the results to the user.

GKB stands for Geographic Knowledge Base. It is a repository of geographic data and knowledge rules relating the data [Chaves et al., 2005]. GKB is used by the Tumba! search engine to formulate queries and assess the geographic context of queries and documents.

4.5 Sidra5 in the XLDB GeoCLEF prototype

The XLDB GeoCLEF framework is a prototype developed for the 2007 edition of GeoCLEF[**?**]. It was used to evaluate the search results provided by the geographic signature approach. The system architecture of this framework is presented in Figure 4.5, and includes the following components in addition to Sidra and GKB, presented above:

QueOnde is responsible for the query interpretation and their segmentation into triplets of: non-geographic terms, geographic relationship, geographic terms [Cardoso et al., 2007].

QuerCol perform query expansion, both textually and geographically, using the triplets generated by *QueOnde* [Cardoso et al., 2007]. The latest version of QuerCol considers both the names and feature type names defined in the geo ontology to adjust the query expansion to the type of geographic entities considered [Cardoso and Silva, 2007].

Faísca is a text mining component, responsible for the extraction and disambiguation of geographic references found on documents. It is this component that generates the geographic signatures of documents [Cardoso et al., 2007].

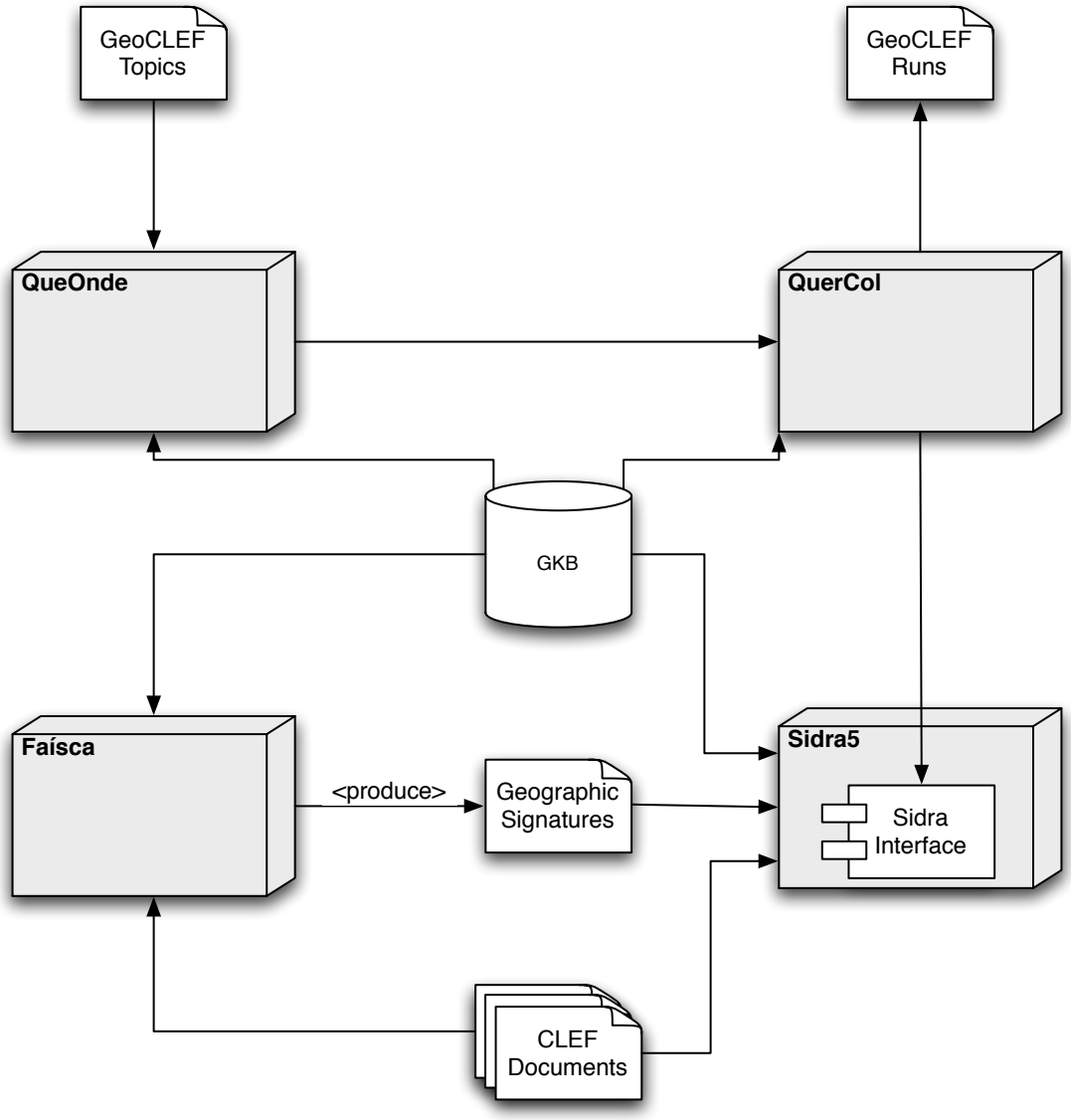


Figure 4.5: XLDB's GeoCLEF evaluation prototype architecture

Sidra5 was fully implemented according to the specified architecture. Section 4.6 describes the selection of the IR framework that was used for the TextIR component.

The evaluation of Sidra5 is done on Chapter 6, using the GeoCLEF prototype as the test vehicule to evaluate the validity of the geographic signatures implementation.

4.6 Text IR Framework

Sidra was redesigned to emphasis maintenance ease and modularity, as described in Section 4.1. Part of the redesign decisions of Sidra were to select an existing IR

framework to be used as the textual indexing, ranking and search component.

This section describes the selection of the text IR framework. In Section 4.6.1, the requirements are presented; in Section 4.6.2, several IR frameworks are presented and evaluated according to the requirements; in Section 4.6.3, some performance tests are specified and in Section 4.6.4, the pre-selected IR frameworks are compared using the performance tests.

4.6.1 Evaluation Criteria

The following criteria were considered in the evaluation of the IR libraries under consideration:

Licensing Terms: the possibility of modifying, adding new features and redistribution of the software are critical.

Price: an essentially relevant issue for non-profit organizations, such as university research groups. Free software is preferred.

Documentation: availability is crucial for a seamless and fast integration of the IR library into the existing system as well as in the development of additional features to add to the IR library.

Implementation Language: may be crucial since it may prevent integration with already existing components. Since the other existing modules are written in Java, software written in Java or with Java bindings is highly preferred.

Operating System: some libraries may work only on some platforms, making them unsuitable for the current deployment strategy. Linux support is mandatory.

Distributed indexes: search engines are bound to very strict performance requirements, which lead to the division of the computing tasks among several computers. To prevent bottlenecks, distributed indexes that can be generated in parallel are indispensable.

Active development: in software tends to be improved more often and new features, such as ranking algorithms, are integrated faster. Also, actively developed software has usually an user base created around it, which helps in getting support.

BM25: Okapi's BM25 ranking algorithm is considered as the state-of-the-art probabilistic weighting scheme and baseline for many IR experiments[Robertson et al., 1995]. Its use can be important for obtaining a good search quality.

Anchor indexing: navigation between web pages is made using links that connect them. The pages targeted by the links are described using a short text snippet, called anchor. For a system which indexes web pages, it is essential that, during the indexing phase, the anchor text is indexed together with the page pointed by the link.

Plug-in architecture: Software libraries with a plug-in architecture have greater extensibility. It is possible to add a functionality without touching the core code of the library, thus easing the implementation of new features with minimum impact to the existing code base.

4.6.2 Evaluated IR Frameworks

In the process of choosing a substitute for Sidra4, several IR frameworks were considered, of which two were selected to be further evaluated, due to time restrictions. The initial phase selection was conducted by an initial mix of feature requirement matching (see Table 4.1) and by some informal testing. The two frameworks selected for further testing were:

Apache Nutch: an open-source search engine implemented in Java. It uses Apache Lucene as its indexing and search component, has an integrated web crawler, a MapReduce facility and support for a distributed file system through the use of Apache Hadoop[3]. Nutch is a very efficient search engine, although lacking some state-of-the-art ranking algorithms, such as Okapi's BM25. It has created a huge community around it providing support and documentation. One of its key features is the ability to extend its functionalities through the use of self contained software plug-ins. Nutch is developed by the Apache Foundation. For more information, see `http://lucene.apache.org/nutch/`.

MG4J is an open-source search engine implemented in Java by the University of Milan. The main objective of MG4J is to offer a highly efficient framework, providing state-of-the-art ranking algorithms, for building search systems. MG4J provides a whole arsenal of functionalities, from distributed processing and index clustering, flexible index construction, to powerful search operators. More than an easy to use framework, MG4J provides full control over the indexing, ranking and search process. For more information, see `http://mg4j.dsi.unimi.it/`.

The following IR frameworks were initially considered for inclusion into Sidra5 as textual indexing and ranking module, but were considered unsuited:

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

Feature	**Sidra4**	**Nutch**	**MG4J**	**Lucene**	**Egothor**	**Lemur**	**Terrier**	**Zettair**
Licence	internal	APL	LGPL	APL	BSD	BSD	MPL	BSD
Price	free	free	free	free	free	free	free	free
Documentation	scarce	abundant	moderate	abundant	scarce	abundant	moderate	scarce
Implementation Language	C++/Java	Java	Java	Java	Java	C/C++	Java	C
Operative System	Linux	Java VM	Java VM	Java VM	Java VM	Linux/Win32/...	Java VM	Linux/Win32/...
Distributed indexes	yes	yes	yes	yes	no	yes	no	yes
Active development	no	yes	yes	yes	no	yes	yes	yes
BM25	yes	no	yes	no	no	yes	yes	yes
Anchor indexing	yes	yes	yes	no	yes	yes	yes	yes
Plug-in architecture	no	yes	no	no	no	no	no	no

Table 4.1: Feature matrix of several IR libraries

Apache Lucene is a high-performance, full-featured text search indexing and searching library written entirely in Java. Apache Lucene is highly reputed for its performance and scalability, and is vastly used worldwide. There is abundant documentation, including books describing its features [Erik Hatcher, 2004]. Lucene is developed by the Apache Foundation.

Despite its qualities, Lucene was not selected for further evaluation, because it lacks the capabilities to parse HTML files and to correctly index/search the information contained on anchors. See `http://lucene.apache.org/java/docs/index.html` for more information.

Egothor is an open source search engine implementation written entirely in Java to ensure cross platform compability. It supports many common file formats, such as HTML, PDF, PS, DOC, XLS. Its architecture enables easy provision for additional file formats. Egothor supports both Boolean and vector searches.

However, scalability is not the main focus of this search engine, being currently used mostly as a demo or in small scale projects. See `http://www.egothor.org/` for more information.

Lemur Toolkit 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++, unlike the software used and implemented both in Tumba! and GeoCLEF prototypes. This differentiation from the current software ecosystem would make its integration more difficult. See `http://www.lemurproject.org/` for more informations.

Terrier is a modular platform for the rapid development of large-scale IR applications, providing indexing and retrieval functionalities, developed by the Information Retrieval Reasearch 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.

Despite all the appealing features of Terrier, only the internal version has all these features. The publicly available software is a crippled version, which lacks several of the cutting-edge features and cannot work in a distributed setting. See `http://ir.dcs.gla.ac.uk/terrier/` for more information.

Zettair is a compact and fast text search engine designed and written by the Search Engine Group at the RMIT University, Australia. It has been designed for simplicity as well as speed and flexibility, being the handling of large amounts of text one of its primary features.

Yet, its irregular release, lack of updates (the last published version is from 8th September 2006) combined with its implementation in C undermine the efficiency and scalability provided by this software. See `http://www.seg.rmit.edu.au/zettair/` for more informations.

4.6.3 Test Specifications

The comparative performance evaluation between Nutch and MG4J was conducted according to three criteria: search quality, indexing time and index size. The TREC GOV1 collection was used as the collection of documents for these tests[**?**]. The GOV1 collection, as subset of the Internet, composed of US government websites, has been shown to be sufficiently representative of the internet and it is characterized by a good proportion of inbound/outbound links and a high connectivity between documents [Soboroff, 2002].

Objectives	evaluate the quality of the results retrieved by the IR framework
Collection	1,2 millions of documents from the GOV1 collection of TREC
Hardware	Irrelevant
Measurements	Mean Average Precision (MAP)

Table 4.2: Search quality test specification

Search Quality: Search quality is vital for the success of any search system. The choice of the IR framework is determinant for the geographic search quality that is built on top of the IR component on Sidra. To help in this selection, the test ecosystem of TREC GOV1 was used (see Table 4.2).

Indexing Time: For large collections of documents, such as the Portuguese web, which is the target of the Tumba! search engine, the indexing time becomes a crucial factor, due to the lack of computational resources. This test intends to find which of the IR frameworks is better suited for large-scale indexing. For the test specification, see Table 4.3. The *Link inverting* measurement consists in the time spent, by the indexeres, to extract the text anchors and associate to document that it refers. The *indexing time* measures consists in the time spent by the indexer to create the index batches and combine them as the final index.

Objectives	measure the indexing time of the IR frameworks
Collection	1,2 millions of documents from the GOV1 collection of TREC
Hardware	CPU: 2x Intel Xeon 3.2GHz (4x virtual CPU) RAM: $\tilde{4}$GB Filesystem: 565GB in Ext3
OS	Version: Red Heat Entreprise Linux 3 Kernel: Linux 2.4.21-47.0.1.ELsmp
Java VM	Version: 1.5.0_11-b03 / options: -Xmx 1024M
Measurements	Links inverting time (in minutes) Indexing time (in minutes) Total time (in minutes)

Table 4.3: Indexing time test specification

Index Size: For large collections of documents, such as the Portuguese web, target of the Tumba! search engine, the size of the indexes is an important factor due to memory costs. For the test specification, see Table 4.4.

Objectives	measure the indexing size obtained by each IR framework
Collection	index of the 1,2 millions of documents from the GOV1 collection of TREC
Hardware	irrelevant
Measurements	size (in MB)

Table 4.4: Index size test specification

4.6.4 Evaluation Results

Search Quality: MG4J obtained better results, showing a higher average MAP value than Nutch (0.3662 versus 0.3168). For each topic, both search systems have shown similar trends, with MG4J being more consistent. It had fewer results with MAP values of zero, meaning that it as an higher overall recall (see Table 4.5). The interpolated-precision/recall curve shown in Figure 4.7 further shows the superiority of MG4J.

Indexing Time: The results of this tests clearly shows the performance edge that MG4J has over Nutch. Despite the time similarity in the "invert links" task (MG4J was faster by 7.9%), the diference in indexing time is abyssmal with MG4J requiring only 56.3% of the time needed by Nutch for the same task.

Indexes Size: MG4J is the winner in this test (see Table 4.6). It was able to outperform Nutch, on index size, by 0.065%, which corresponds to a difference

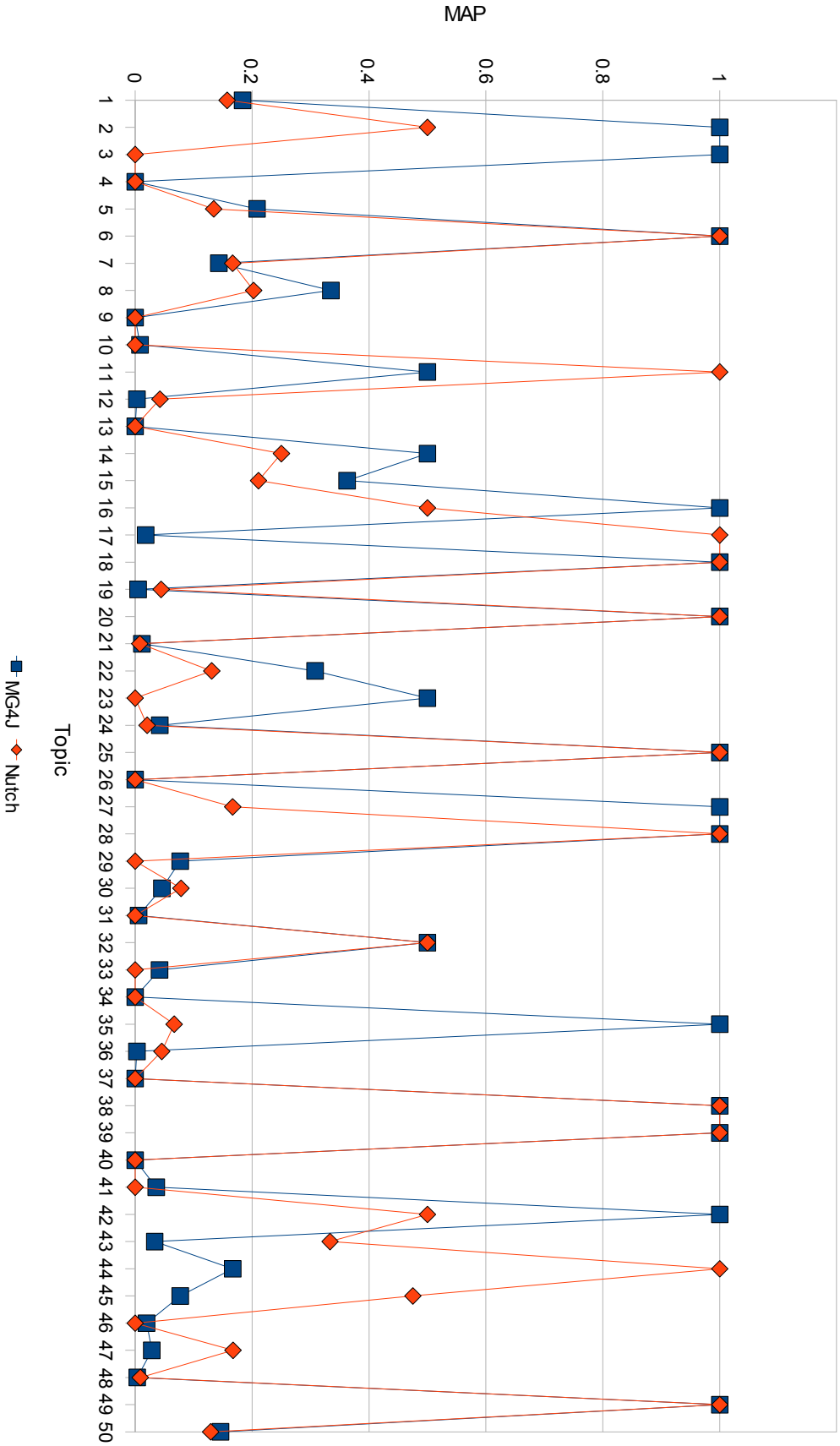


Figure 4.6: MAP values for the 50 TREC topics on the Gov1 collection

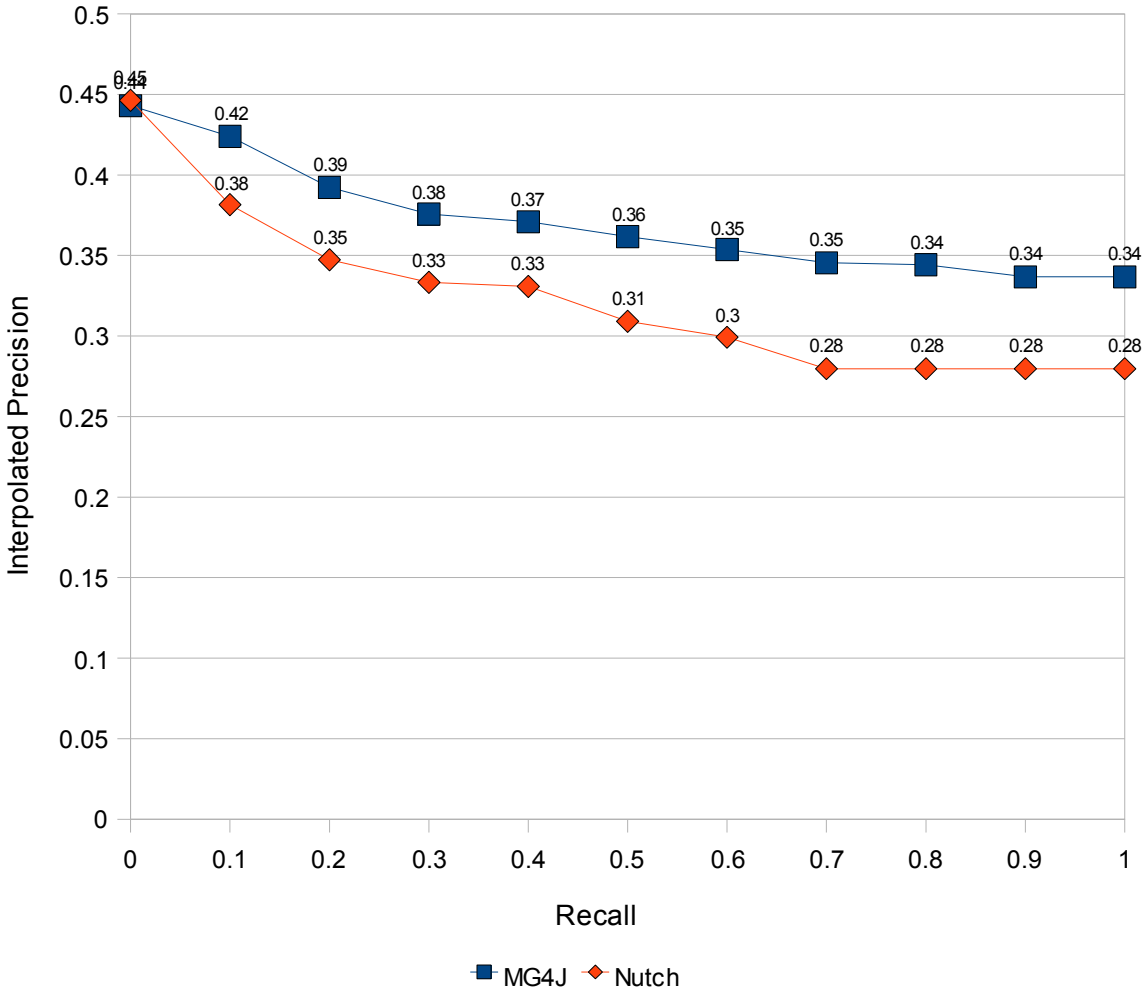


Figure 4.7: Interpolated-precision/recall curve for the TREC results

of almost 200MB in indexing a collection of 5847MB (compressed in GZip format). The difference between both frameworks may seem small for today's standards, but the 1.2 million documents of GOV1 are nowhere near the size of a small partition of the internet like the Portuguese web.

Superior performance combined with better search algorithms make MG4J the best choice as text retrieval component of Sidra5. It has all the main requirements: it is free, open-source, and in active development, it runs in Linux, indexes HTML anchors, implements Okapi BM25 and it is written in Java. Given the above features, it has shown to be easily integrated into the existing software infrastructure of Tumba! and also easily embeddable in the GeoCLEF prototype.

This chapter presented the architecture of Sidra5 and the structure of its indexes; showed how to integrate Sidra into both the XLDB GeoCLEF prototype and the Tumba! search engine (the integration into Tumba! was not effectuated); tested several text IR frameworks and selected one of them to be included into Sidra. The

Table 4.5: MAP values for the TREC topics

Topic	MG4J	Nutch	Run	MG4J	Nutch
1	0.1836	0.1573	26	0.0000	0.0000
2	1.0000	0.5000	27	1.0000	0.1667
3	1.0000	0.0000	28	1.0000	1.0000
4	0.0000	0.0000	29	0.0769	0.0000
5	0.2085	0.1343	30	0.0458	0.0784
6	1.0000	1.0000	31	0.0058	0.0000
7	0.1429	0.1667	32	0.5000	0.5000
8	0.3348	0.2026	33	0.0417	0.0000
9	0.0000	0.0000	34	0.0000	0.0000
10	0.0083	0.0000	35	1.0000	0.0667
11	0.5000	1.0000	36	0.0028	0.0455
12	0.0028	0.0424	37	0.0000	0.0000
13	0.0000	0.0000	38	1.0000	1.0000
14	0.5000	0.2500	39	1.0000	1.0000
15	0.3626	0.2109	40	0.0000	0.0000
16	1.0000	0.5000	41	0.0359	0.0000
17	0.0179	1.0000	42	1.0000	0.5000
18	1.0000	1.0000	43	0.0333	0.3333
19	0.0049	0.0444	44	0.1667	1.0000
20	1.0000	1.0000	45	0.0771	0.4750
21	0.0114	0.0079	46	0.0196	0.0000
22	0.3076	0.1310	47	0.0284	0.1672
23	0.5000	0.0000	48	0.0036	0.0086
24	0.0420	0.0204	49	1.0000	1.0000
25	1.0000	1.0000	50	0.1460	0.1290
Average:				0.3662	0.3168

	MG4J	Nutch	Variation (in %)
Links inverting (min)	11.03	11.98	7.9%
Indexing time (min)	64.35	147.12	56.3%
Index size (MB)	2768	2960	0.065%

Table 4.6: Comparison of indexing time and index size between MG4J and Apache Nutch

next chapter will detail how geographic signatures can be used to rank documents according to their relevance.

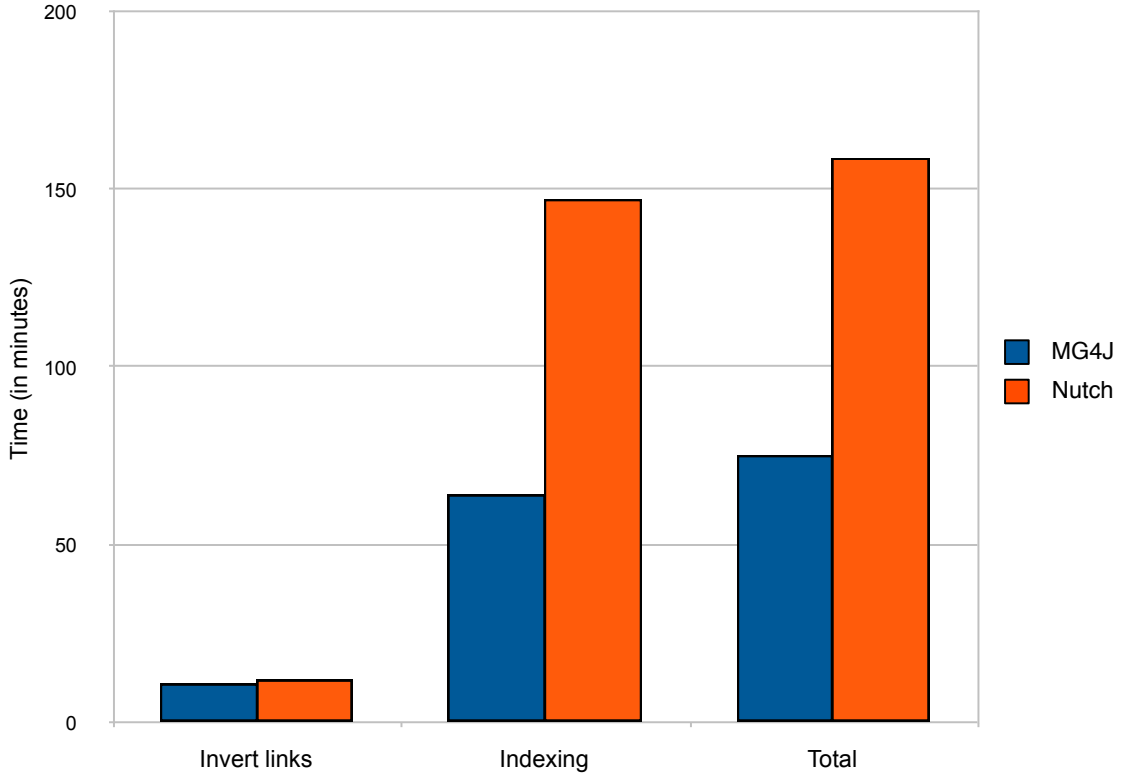


Figure 4.8: Indexing time

Chapter 5

Ranking with Geographic Signatures

A query is geographic when it contains geographic information. A good query analysis is crucial for the interpretation of the query context, correctly identifying the existing geographic terms or references.

In the used GIR architecture, the input query is submitted to a component responsible for the geographic information extraction. As shown in the previsou chapter, this component is either the *Presentation Engine* of Tumba! or QueOnde in the GeoCLEF prototype. From now on, the description will focus on the search process from the GeoCLEF prototype point of view.

The search process is briefly described in Figure 5.1. The identification of the query structure is essential to distinguish geographic queries from non-geographic ones. *Sidra* uses the query structure to drive the retrieval and ranking processes. If the geographic part is present, it does a geographic retrieval, else it does a classic textual retrieval.

Independently of the query type, both search types share the common steps of textual retrieval:

1. Fetch from the textual index the identifiers of the documents (docid) that match the thematic part of the query

2. Score the documents using the text ranking algorithms

3. Rank the documents

If the query is purely textual, the search ends with the selection of the result sub-set from the top ranked documents. If the query is geographic, some additional steps have to be taken:

1. Fetch from the geographic index the geographic signatures of the documents

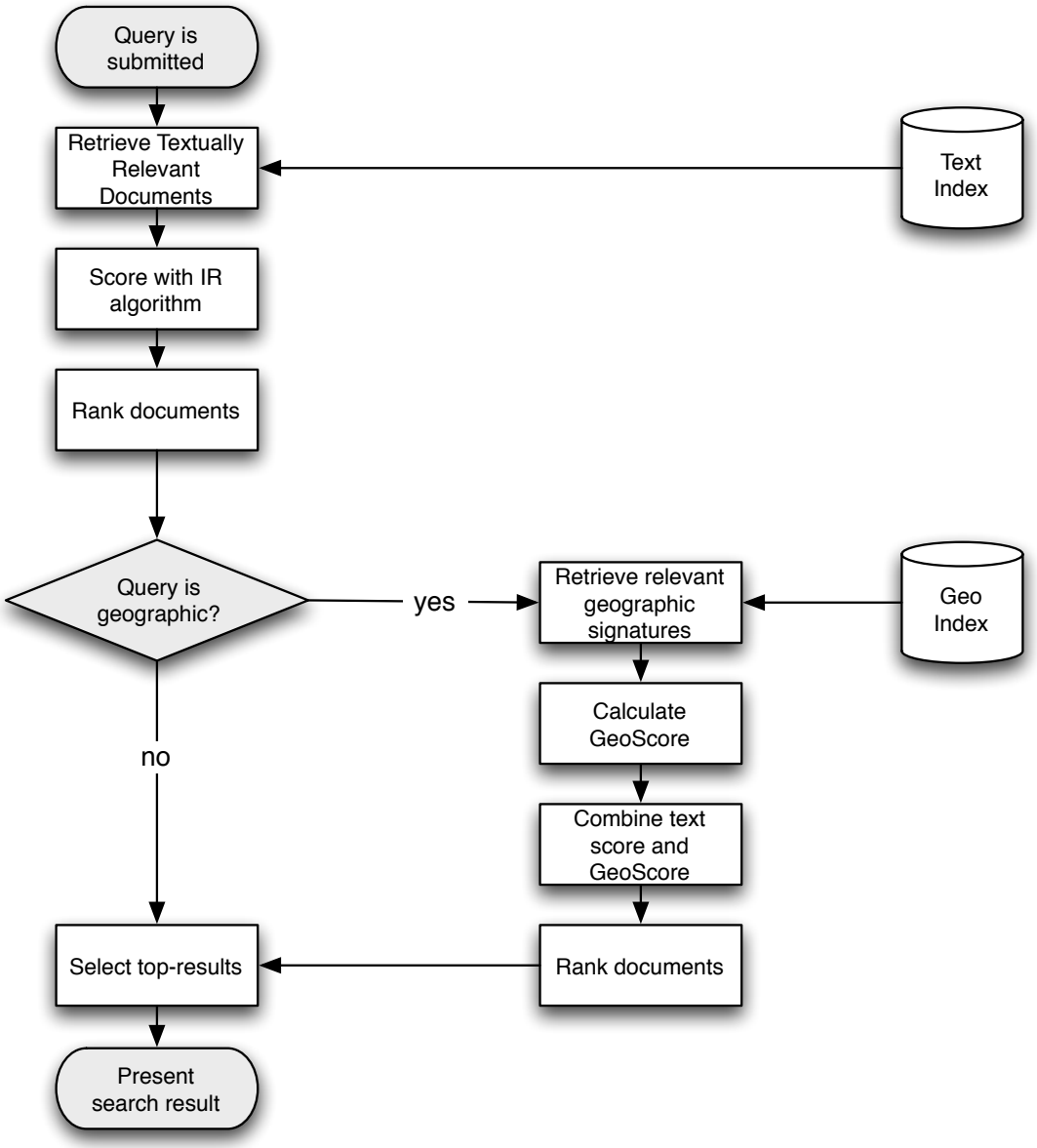


Figure 5.1: Flow chart of searches in Sidra5

2. Calculate the similarity score between the geographic signatures of the query and the documents using one of the implemented geographic similarity strategies

3. Combine the textual and geographic score of the documents into one unified score

4. Rank the documents from higher to lower score

The rest of this chapter present the structure of the geographic signatures (Section 5.1, explains how the geographic signatures are used to geographically rank

documents and presents the heuristics and strategies used to compare between two geographic signatures (Section 5.2).

5.1 Geographic Signature Definition

Sidra5 uses the geographic signature as the data structure for the geographic indexing, ranking and search.

Almost all the geographic reasoning and processes in Tumba! and in the GeoCLEF prototypes, presented in the last chapter, revolve around the representation of geographic information in data structures called *geographic signatures*. The geographic signature approach considers that the geographic context of a document is represented by multiple geographic references, in a way similar to Markowetz et al. [2005] and mainly the probabilistic approach of Yi Li and Cavedon [2006].

Geographic signatures represent the geographic information present in queries and documents. A geographic signature is a list of identifiers of geographic features present in the ontology and their corresponding confidence measure ($ConfMeas$), a value in the [0,1] range, which represents the confidence that the feature is part of the geographic scope. $ConfMeas$ is obtained through an analysis of the surrounding concepts on each case, in a similar way as described by Yi Li and Cavedon [2006].

In the current implementation of the system, geographic signatures of queries (Q_{sig}) have $ConfMeas$ set to 1.0.

Geographic signatures of documents (D_{sig}) are generated by extracting geographic names from documents. An excerpt of four document signatures (one per line) from the GeoCLEF collection is given below:

```
FSP940101-023:  1432[1.00]
FSP940101-024:  6379[1.00]; 6380[1.00]; 6381[1.00]; 8737[1.00]
FSP940101-025:  5838[1.00]; 12332[0.67]; 1048[0.33]; 89[0.33]; 1889[0.33]
FSP940101-026:  5838[1.00]; -14[0.25]; 12332[0.25]; 5917[ 0.25]; 8734[0.25]
```

Despite being very similar in structure, query and document signatures represent different levels of geographic information. A Query Geographic Signature (Q_{sig}) is a list of identifiers of geographic features present in the used ontology. The purpose of Q_{sig} is to enumerate all the geographic references present in the query for later search and document similarity evaluation.

5.2 Geographic Ranking

In IR, document ranking means ordering the documents by decreasing relevance relatively to the query terms. Geographic ranking does not change the foundation

of this principle, but requires new document ranking algorithms. The document relevance calculation of this work is based on the previous work by Martins et al. [2006], adapted to support the processing of geographic signatures.

Because geographic signatures comprise multiple geographic features, an additional processing layer was added to the ranking process, and several geographic scoring strategies that could be employed to evaluate the level of resemblance between geographic signatures have been compared. Algorithm 1 represents the generic ranking framework, independent of the geographic scoring method employed.

Algorithm 1 Generic Ranking Algorithm

Require: relevant_documents
Require: Q_{sig}
Require: txt_weight
  for all doc in $relevant\_documents$ **do**
    $tmp_score \leftarrow txt\_weight \times doc.score$
    $D_{sig} \leftarrow getD_{sig}(doc.id)$
    for all s_{doc} in D_{sig} **do**
      for all s_{query} in Q_{Sig} **do**
        $geoScore \leftarrow computeGeoScore(s_{query}, s_{doc})$
        $tmp_score \leftarrow updateScore(geoScore, 1 - txt\_weight)$
      end for
    end for
    $doc.score \leftarrow tmp_score$
  end for

The final score of a document, for a given query, consists in the linear combination of the document textual relevance score and the geographic score between Q_{sig} and D_{sig}. The textual score is obtained when the textual part of the query is evaluated using the Okapi BM25 ranking algorithm [Robertson et al., 1995], normalized to values in the interval [0,1] as defined by Song et al. [2004]. The geographic score ($GeoScore$) is obtained by calculating the geographic similarity between the n geographic features of the Q_{sig} with the m geographic features present in the D_{sig} using some evaluation strategy. The final score is obtained using the following formula:

$$\begin{aligned} Ranking(query, doc) = \; & \mathbf{TxtWeight} \times NormBM25(query, doc) \; + \\ & \mathbf{GeoWeight} \times GeoScore(query, doc) \end{aligned} \tag{5.1}$$

The constants $TxtWeight$ and $GeoWeight$ were both set to 0.5.

For this work, four $GeoScore$ strategies were created:

Maximum: Considers that the value of $GeoScore$ between Q_{sig} and D_{sig} is the maximum similarity value between one geographic feature from Q_{sig} and one

from D_{sig}. This evaluation strategy is represented by the following formula:

$$GeoScore_{max}(query, doc) = max(GeoSim(s_1, s_2) \times ConfMeas(s_2)) \ , \ s_1 \in Q_{sig} \wedge s_2 \in D_{sig} \tag{5.2}$$

Mean: Considers that the *GeoScore* is the average of the similarity values of combining Q_{sig} geographic features with D_{sig} geographic features. This evaluation strategy is represented by the following formula:

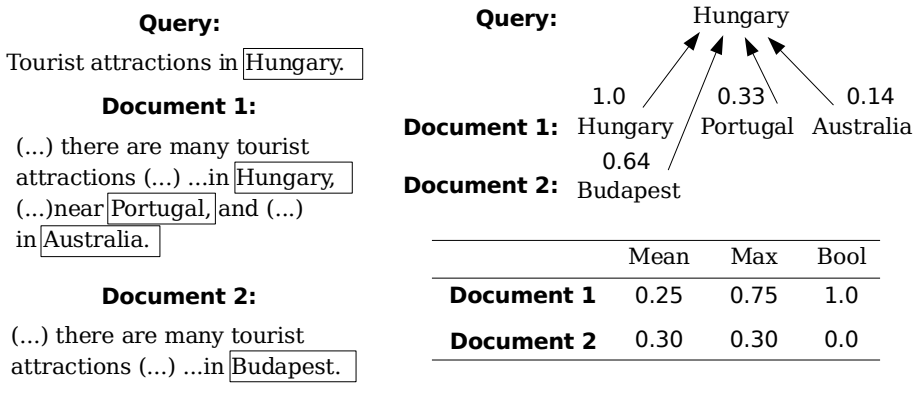
$$GeoScore_{mean}(query, doc) = avg(GeoSim(s_1, s_2) \times ConfMeas(s_2)) \ , \ s_1 \in Q_{sig} \wedge s_2 \in D_{sig} \tag{5.3}$$

Boolean: Considers that the *GeoScore* value is 1 if D_{sig} contains any geographic feature that is also present in Q_{sig}, and zero otherwise. This evaluation strategy is represented by the following formula:

$$GeoScore_{bool}(query, doc) = \begin{cases} 1 & \text{if } \exists \ s_1 = s_2, \ s_1 \in Q_{sig} \wedge s_2 \in D_{sig} \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

Null: $GeoScore_{null}$ is always zero, turning off the geographic scores. This is used as a baseline metric for comparing results obtained with the other metrics.

In the formulae, *ConfMeas* corresponds to confidence measure associated to each geographic feature of D_{sig}, as explained in Section 5.1.



	Mean	Max	Bool
Document 1	0.25	0.75	1.0
Document 2	0.30	0.30	0.0

Figure 5.2: Example of the calculation of the four GeoScore combination metrics

The computation of the four *GeoScore* metrics is illustrated in Figure 5.2, which presents a fictional query (*Hungary*), and two document surrogates, along with the $GeoSim \times ConfMeas$ values and final *GeoScore* values.

Next, the textual relevance calculation of documents and the geographic similarity calculation are presented in detail.

Textual Ranking

The retrieval of potentially relevant documents to a query uses the textual index. To these documents is then applied a ranking algorithm that assigns a relevance score to each one and orders them by relevance order. This textual ranking is calculated by the probabilistic algorithm BM25, where the relevance of each document consists in the weighted sum of terms that occur in the document and in the query [Robertson et al., 1995]. The weight of each term t_i is given by the following formula:

$$BM25(t_i) = \frac{(k_1+1) \times term\_freq(t_i)}{k_1 \times ((1-b) + b \times \frac{doc\_len}{avg\_doc\_len}) + d} \times \log(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5}) \quad (5.5)$$

The constants k_1 and b were set to the default values of 2.0 and 0.75, respectively.

Because the results obtained by BM25 aren't necessarily in the interval [0,1], needed for the final ranking score formula, it is necessary to normalize the returned scores. The used normalization formula used is described by Song et al. [2004] and corresponds to:

$$NormBM25(query, doc) = \frac{\sum_{t_i \in doc} BM25(t_i) \times weight(query, t_i)}{\sum_{t_i \in doc} \log(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5})(k_1 + 1)} \quad (5.6)$$

Geographic Similarity

Independently of the strategy used for the *GeoScore* computation, each Q_{sig} geographic feature has to be compared to each D_{sig} geographic feature. This process corresponds to geographic similarity calculation and uses four heuristics: ontological similarity, distance similarity, populational similarity and adjacency similarity. The similarity value is obtained by the following formula:

$$\begin{aligned} GeoSim(s_1, s_2) = &(g_1 \times OntSim(s_1, s_2)) + (g_2 \times DistSim(s_1, s_2)) + \\ &(g_2 \times PopSim(s_1, s_2)) + (g_4 \times AdjSim(s_1, s_2)) \end{aligned} \quad (5.7)$$

where the constants g_1 through g_4 have been set to:

- $g_1 = 0.5$

- $g_2 = 0.2$
- $g_3 = 0.2$
- $g_4 = 0.1$

Next, the used geographic heuristics are presented.

Ontological Similarity: Topological Relations of the type "part-of" defined on the ontology, can be used to infer similarity degrees. This similarity is calculated using the number of transitions between common ancestors. The formula used, similar to the measure of Lin [1998], is:

$$OntSim(scope_1, scope_2) = \begin{cases} 1 \text{ if } scope_1 \text{ is equal or equivalent to } scope_2 \\ \frac{2 \times NumCommonAncestors(scope_1, scope_2}{NumAncestors(scope_1) + NumAncestors(scope_2)} \end{cases} \text{otherwise} \quad (5.8)$$

Distance Similarity: There is an implicit notion that near locations are more similar than two more distant locations. However, this distance notion is context-dependent, varying with the size and reference point between both locations.

In this measure, a similarity value of one is reached when some geographic entity is contained inside another or, in other words, the distance between them is zero. A double sigmoidal function is used with a center corresponding to the *minimum bounding rectangle* diagonal. This function has maximal value when distance is minimal, from which it declines slowly to zero as distance augments, providing a non-linear normalization. In the next formula, D corresponds to distance between $scope_1$ and $scope_2$ and where D_{MBR} corresponds to the *minimum bounding rectangle* diagonal distance corresponding to $scope_2$.

$$DistSim(scope_1, scope_2) = \begin{cases} 1 \text{ se } scope_1 \text{ is included in or is father of } scope_2 \\ 1 - \left(\frac{1 + sign(D - D_{MBR}) \times (1 - \exp(-(\frac{D - D_{MBR}}{D_{MBR} \times 0.5})^2))}{2}\right) \end{cases} \text{otherwise} \quad (5.9)$$

Populational Similarity: When an inclusion relationship exists between two geographic features, a partial "part-of" relation, the fraction obtained from the comparison of the population of the more specific area with the more global area can be used as a similarity value. This measure shows the relative im-

portance that a region as inside another, and is obtained by:

$$PopSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is equivalent to } scope_2 \\ \frac{PopulationCount(scope_1)}{PopulationCount(scope_2)} & \text{is } scope_1 \text{ included in } scope_2 \\ \frac{PopulationCount(scope_2)}{PopulationCount(scope_1)} & \text{is } scope_2 \text{ included in } scope_1 \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

Adjacency Similarity: Considering that two adjacent locations are more relevant than two that aren't, the information provided by the ontology about adjacency between geographic features is used. A value of one is assigned if two features are adjacent and zero otherwise:

$$AdjSim(scope_1, scope_2) = \begin{cases} 1 & \text{if } scope_1 \text{ is adjacent to } scope_2 \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

Geographic signatures are a novel approach in the modeling of geographic information for GIR purpose.

In this chapter, the processes needed for geographic search with signatures was described. An indexing scheme was presented which allows efficient and flexible storage of textual and geographic information. A geographic ranking scheme was presented which uses the geographic signatures for the relevance reasoning, describing several distinct geographic scoring strategies that use the heuristics described in Martins et al. [2006].

In the next chapter, the geographic signature approach is evaluated. Test scenarios are presented and the obtained results are evaluated and compared with a pure textual approach.

Chapter 6

Evaluation

This section presents the evaluation of the geographic signature implementation of Sidra5 using the XLDB GeoCLEF test infrastructure. GeoCLEF was choosen because of being a forum devoted to GIR, and having geographic tasks in Portuguese (the target of the Tumba! search engine) and English (the more commonly used language in the internet), among others languages[**?**].

This chapter starts with Section 6.1, specifies tests with the purpose of identifying the best GeoScore, the impact of several query expansion strategies and to observe how geographic signatures compare to text retrieval. Secondly, Section 6.2 present and interpret test results.

6.1 Specification

The purpose of the tests conducted was to evaluate how the use of geographic semantics contributes to a bigger query contextualization and improved search results, validating the geographic signatures centred approach.

The document collection used for the Sidra5 evaluation was taken from the CLEF document collection. It is composed by the articles of 1994 and 1995 of the newspapers *Público* and *Folha de São Paulo*, for the Portuguese tests, and by the articles of 1994 of the newspaper *Los Angeles Times* and the articles of 1995 of the newspaper *Glasgow Herald*, for the English tests (for more details see Tables 6.1a and 6.1b). As part of the indexing process of these collections, the D_{sig} for both the Portuguese and English documents were produced. Figures 6.1a and 6.1b represent the distribution of D_{sign} by geographic signature size, i.e. the number of geographic feature references present, and Table 6.2 present some statistics about the D_{sig} distribution.

Sidra5 was tested using the 25 English search topics (see Table A.4) and the 25 Portuguese search topics (see Table A.3) of the 2007 edition of CLEF geographic track (GeoCLEF). The results returned from geographic search systems are evaluated through a set of relevance judgments that indicate, for each query, the rele-

Newspaper	Number of documents
Folha de São Paulo 94	51875
Folha de São Paulo 95	52038
Público 94	51751
Público 95	55070
Total:	210734

(a) Portuguese collection

Newspaper	Number of documents
Los Angeles Times 94	113006
Glasgow Herald 95	56472
Total:	169478

(b) English collection

Table 6.1: CLEF document collections

Measures	Portuguese Collection	English Collection
# of D_{sig}	210734	169475
# of geo feature ids	1159039	1476839
Average # of geo feature	5.500	8.714
Mode	0	0
σ	5745.615	3220.322

Table 6.2: Statistics of document geographic signtures

vant documents. The relevance judgments are created through a technique called "polling" [Harman, 1995]. In this technique, a team of evaluators manually judge, for each query, from the top n results of all search systems, which documents are relevant and which are not, and provide information such as the *Mean Average Precision* (MAP), and precision and recall charts.

For the two sets of GeoCLEF documents, four evaluations were made (described in Figure 6.2 and summarized in Table 6.3), covering the different GeoSim strategies that have been implemented:

Terms Only: In this test, geographic expansion (Geo. QE) of geographic terms is made, in other words additional geographic terms are added to the query using the hierarchical information present on the ontology, before a query expansion by *Blind Relevance Feedback* (BRF) [Rocchio Jr., 1971]. In BRF, the top-n documents of an initial query are assumed as being relevant and are used to enhance the following queries.

The particularity of this evaluation is that only terms are used, since the geographic expansion uses the geographic terms instead of the identifiers of the geographic features.

Geo. QE → BRF: In this test, both textual and geographic ranking are used for the queries. A geographic expansion is performed before the expansion of the thematic part of the query by BRF.

BRF → Geo. QE: In this test, both textual and geographic ranking are used for the queries. An expansion by BRF of the thematic part of the query is performed before the expansion of the geographic part.

Terms/GIR: This is a hybrid of the previous examples. In a first step, a geographic expansion using terms is performed as a purely textual query, which will feed the process of BRF. For the final query, the result of BRF is submitted as a geographic query which will combine textual and geographic relevance of documents.

Test	Description
Terms Only	Geographic expansion by terms before *Blind Relevance Feedback*. Textual search.
Geo. QE → BRF	Geographic Expansion before *Blind Relevance Feedback*. Geographic search.
BRF → Geo. QE	*Blind Relevance Feedback* before geographic expansion. Geographic search.
Terms/GIR	Geographic expansion by terms before *Blind Relevance Feedback*. Geographic search.

Table 6.3: Descriptions of the tests

The test cases were intended to answer the following questions:

- Does the GIR approach based on geographic signatures make it possible to obtain improved results relatively to a classical IR approach?
- Which *GeoScore* gives the best results?
- At what stage does the order of the geographic expansion gives better performance? Before or after terms?
- Should the geographic terms also be included on the thematic part of the query, for the processing by the text retrieval engine? And what is the impact of this on the results?

	GeoScore	Terms only	Geo.QE→BRF	BRF→Geo. QE	Terms/GIR
Initial run		**0.210**	0.126	0.084	**0.210**
Final Run	Maximum	**0,233**	0.122	0.094	0.205
	Mean		0.022	0.021	0.048
	Boolean		**0.135**	**0.125**	**0.268**
	Null		0.093	0.093	0.221

(a) Portuguese monolingual subtask

	GeoScore	Terms only	Geo.QE→BRF	BRF→Geo. QE	Terms/GIR
Initial run		**0,175**	0.086	0.089	**0.175**
Final Run	Maximum	**0.166**	0.093	0.104	**0.218**
	Mean		0.043	0.044	0.044
	Boolean		**0.131**	**0.135**	0.204
	Null		0.081	0.087	0.208

(b) English monolingual subtask

Table 6.4: MAP results obtained for the GeoCLEF task

6.2 Results

The experimental results show that the use of geographic signatures improved results quality. However, the results indicate that neither a pure textual nor a geographical approach yield the best results, but a combination of both. This shows the importance of geographic terms being evaluated as such and not being stripped from the thematic part of the query and considered on the geographic part only. The results of the Portuguese experiments are presented on Table 6.4a and the results of the English experiments are presented on Table 6.4b.

Analyzing the results, it is possible to conclude that:

- The classical IR approach (*Terms Only*) is consistently superior to the pure GIR techniques (Geo. QE → BRF and BRF → Geo. QE) and is only surpassed by a hybrid textual and geographic approach (Terms/GIR). The use of the geographic terms in the initial run is essential for a more precise query narrowing.

- The $GeoScore_{bool}$ is the strategy that globally gives the best MAP values. Because $GeoScore_{bool}$ is more straightforward on assigning the geographic scores to documents, it is not influenced by the quality of the GeoSim heuristics nor by the quality of the geographic ontology, contrary to $GeoScore_{mean}$ and $GeoScore_{max}$. When $GeoScore_{bool}$ scores documents as being geographically relevant, the differentiation among those documents is made using the textual

score that uses the BM25 ranking algorithm, hence the quality of $GeoScore_{bool}$ resides on the textual ranking quality.

- The $GeoScore_{mean}$ obtained the worst results, independently of the used query expansion strategy (see Figure reffig:map-geoscores-pt. The results obtained by the *Mean* scoring metric are even worse than those of $GeoScore_{null}$, meaning that better results are returned using no geographic reasoning than by using the *Mean* reasoning. The deceptioning results show the inadequacy of the GeoSim heuristics and the query drifting due to long D_{sig}, which affect particularly this GeoScore.

- The benefit of early geographic expansion is only visible on the Portuguese experiments improving the initial run (0.126 versus 0.084), meaning that more relevant documents are present on the top-k docs, thus improving the results from the BRF step (see Figures 6.4a and 6.5a). However, there is no benefit on using early geographic expansion on the English experiments. It is then impossible to conclude on overall usefulness of this technique despite the benefices with the Portuguese queries (see Figures 6.4b and 6.4b).

- The use of geographic terms on the thematic part of the query allow better results, fact that can be observed on the *Terms Only* and *Terms/GIR* tests. These experiments had higher MAP values for the initial run compared to the experiments that used geographic expansion through feature identifiers (0.210 versus 0.126 for Portuguese, see Figures 6.4a and 6.5a). Since the experiments with higher final MAP are those that have higher MAP on the initial run, the evidence shows that the quality of the final run is dependent on the initial run quality (see Figure 6.6a).

- Regarding the English experiments, we observe similar trends as in the Portuguese experiments. The slightly lower values are consequence of the quality of the used ontology, which is more complete with Portuguese feature names.

- The GeoSim heuristics presented and used by Sidra5 presented several deficiencies who negatively contributed to the results.

  Ontological Similarity: In the GeoCLEF 2007 evaluation, 25% of the relevant documents contained geographic names that were not in the ontology, and poor results were obtained when handling queries with unknown geographic names. This lack of information had a non-negligible impact on the more heavily weighted GeoSim heuristic.

  Distance Similarity: This measure was heavily crippled by the lack of bounding box coordinates of small geographic entities, such as cities, on the

ontology. The strong dependency of bounding box coordinates combined with a chronical lack of this information on small entities rendered this heuristic measure highly instable. The *Distance Similarity* needs to be changed to be less vulnerable to incomplete information.

Populational Similarity: Disregarding the possible lack of precision of population values, only a small sub-set of geographic entities on the ontology have information about the population. Even only among administrative geographic features, many of them lack this information. Since this measure only gives non-zero values if both geographic features being compared have the population information, the impact of this measure is minimal and possibly causes query drifting.

The objective of the XLDB prototype for the participation in GeoCLEF was to test the validity of the geographic signatures approach, evaluating if it allowed improved results relatively to a purely textual approach. The prototype also aimed at comparing several scoring metrics that have been proposed and comparing distinct query expansion strategies.

Results showed that geographic approaches alone are unable to surpass the traditional textual IR in terms of search results quality. However, a hybrid approach combining the use of geographic terms as both textual and geographical knowledge units, obtained the best results. Hence GIR and IR need each other to improve search quality, IR narrow the query and GIR to increase query contextualization.

Results are also dependent on the performance of the scoring methods, with the naive $GeoScore_{bool}$ obtaining better results than the competing geographic scores in almost all query expansion strategies. The $GeoScore_{max}$ and $GeoScore_{mean}$ scoring methods were crippled by inadequate geographic similarity heuristics and by the quality of the ontology. The results also showed that the GeoSim heuristics need to be completely reevaluated due to their excessive instability with incomplete results. This fact was one of the reasons for the success of $GeoScore_{bool}$. Since $GeoScore_{bool}$ only uses the geographic signatures for the geographic score, it avoids the query driftings provoked by the *GeoSims*.

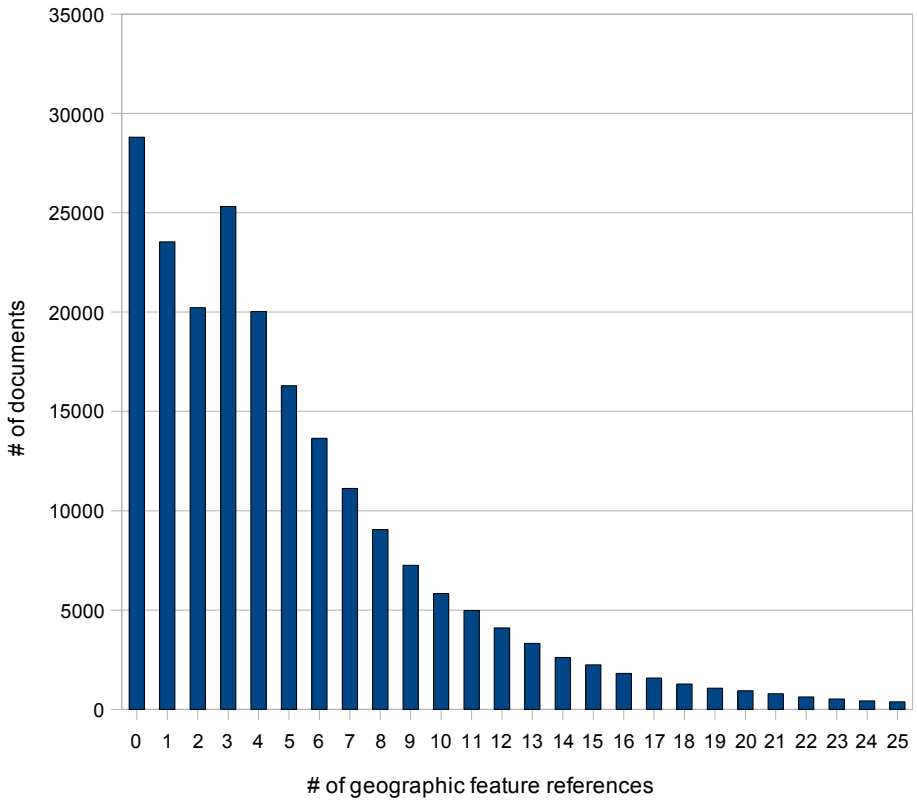
In the GeoCLEF 2007 evaluation, 25% of the relevant documents contained geographic names that were not in the ontology, and poor results were obtained when handling queries with unknown geographic names. In addition, the used ontology is not comprehensive on coordinates and population data to serve the geographic heuristics.

Finally, comparing the Portuguese and English tasks, both presented similar trends. The unique differences worth mentioning are:

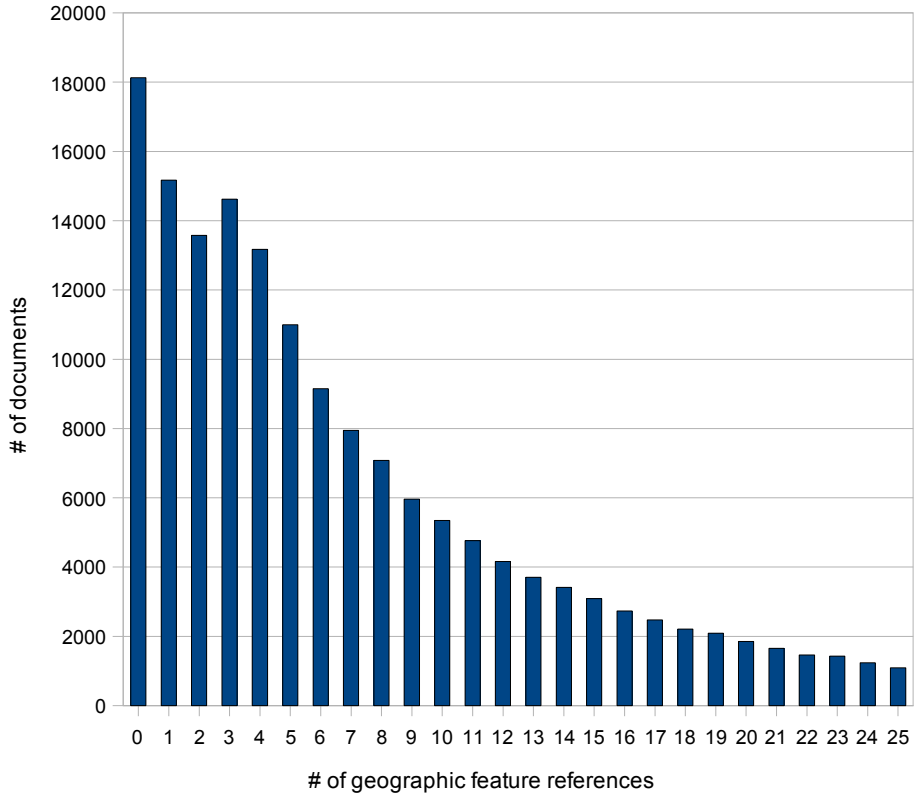
- *Geo. QE* $\rightarrow$ *BRF* was beneficial in the Portuguese task but not in the English task;

- The best result in the English task was obtained by $GeoScore_{max}$ in the Terms/GIR strategy, contrasting with the overall dominance of $GeoScore_{bool}$.

It is safe to assume that the key to improve search results, through the use of geographic signatures, is to use *GeoSim* heuristics and *GeoScores* that are both resistant to errors or missing information and that allow an increasing contexutalization of the geographic reasoning.



(a) Portuguese collection



(b) English collection

Figure 6.1: Histogram of length of geographic signatures in documents

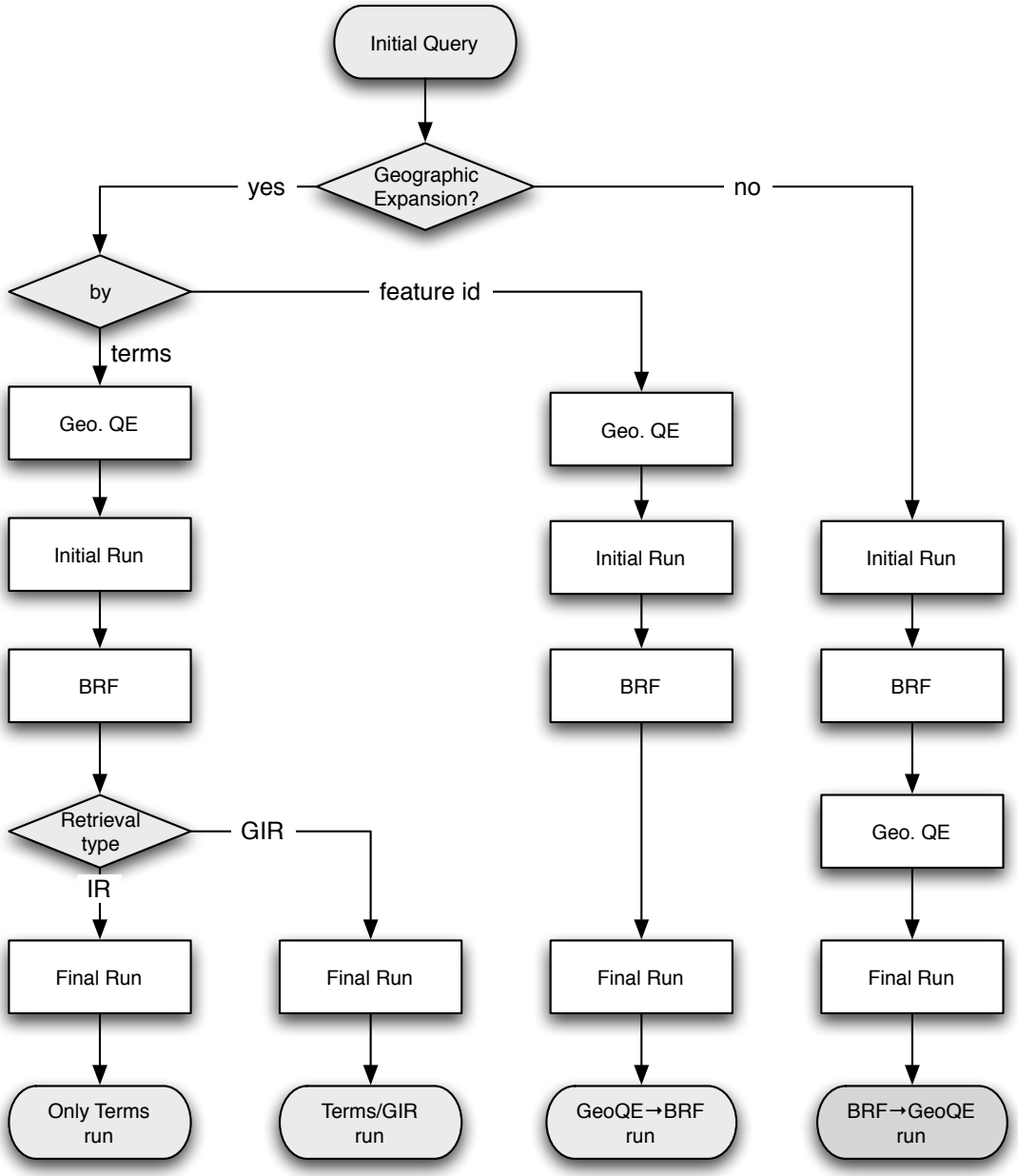


Figure 6.2: Runs flux diagram

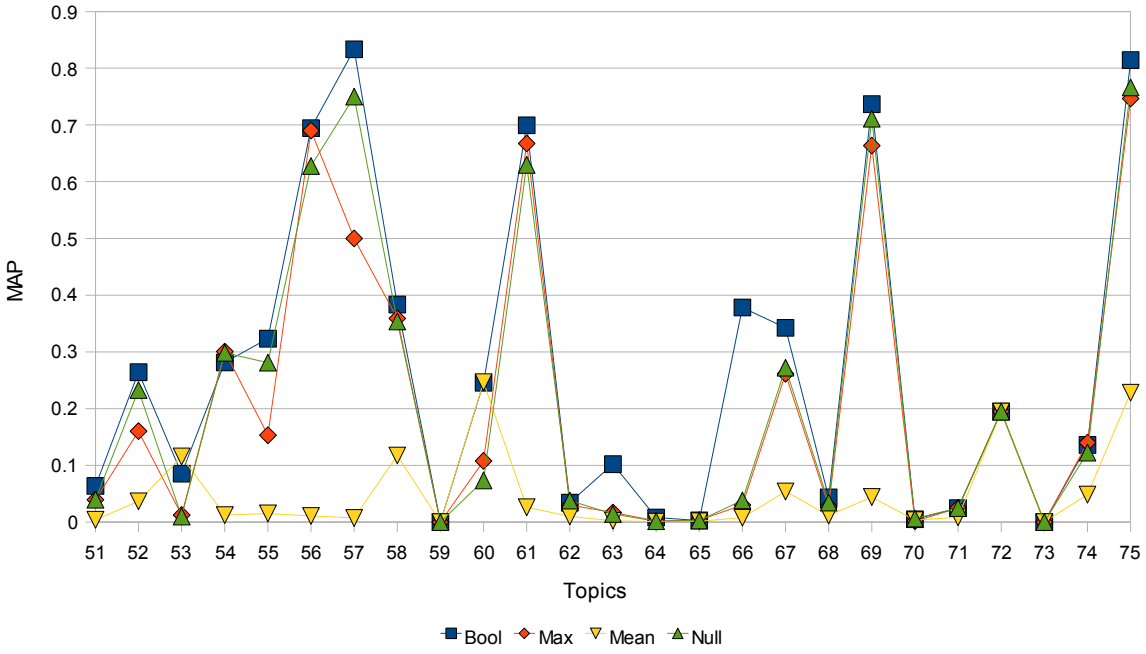
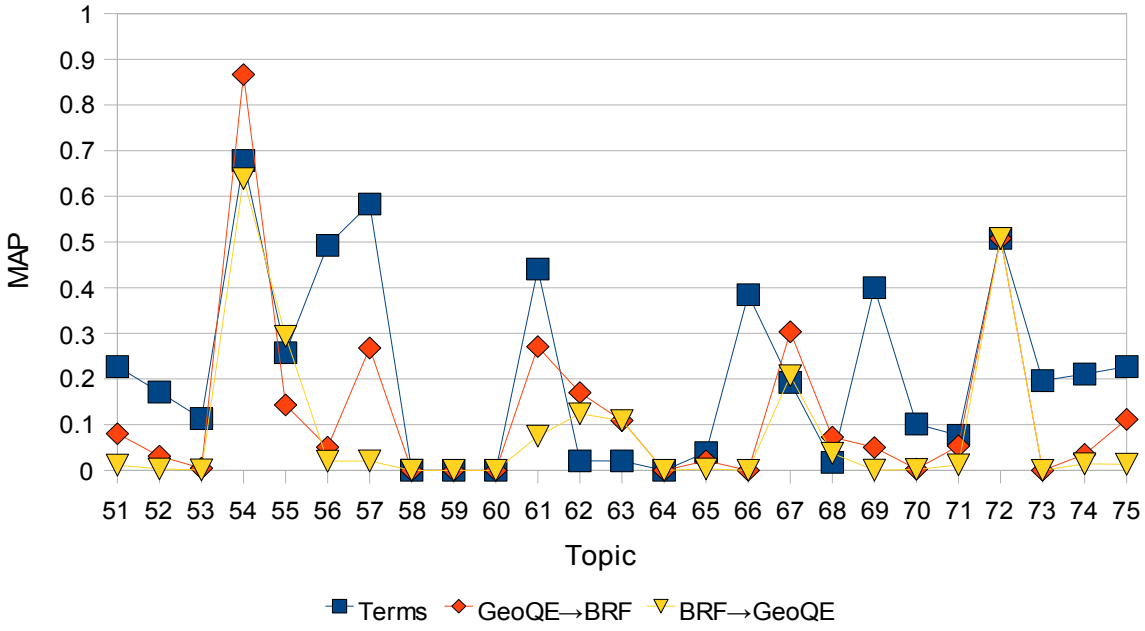
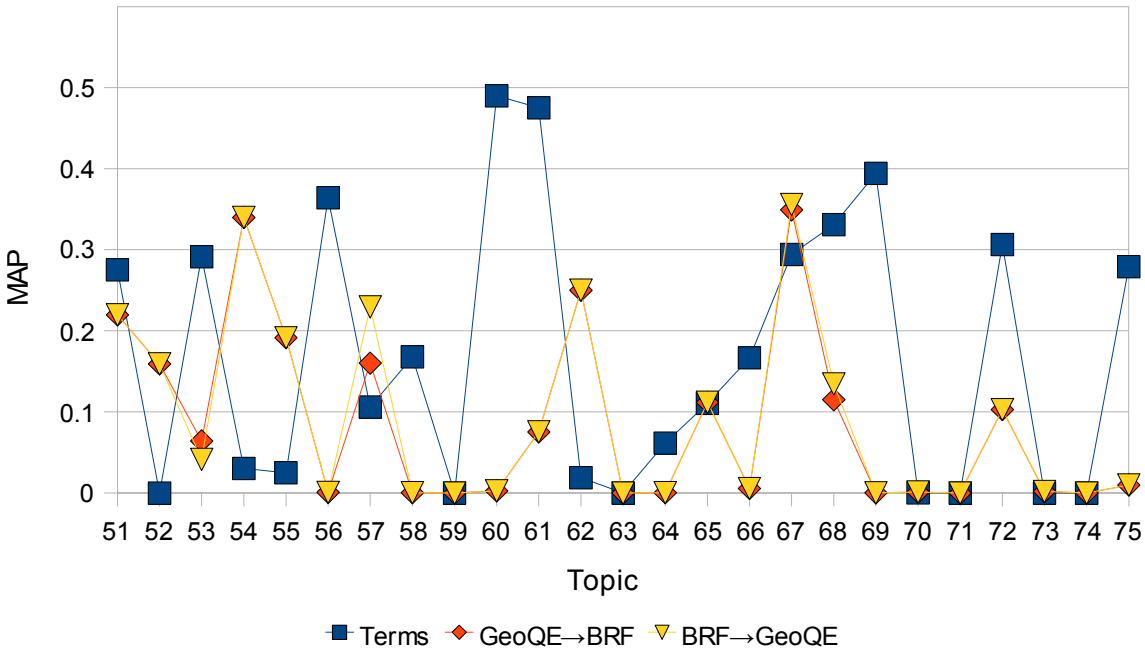


Figure 6.3: MAP values of the GeoScores for the Terms/GIR strategy

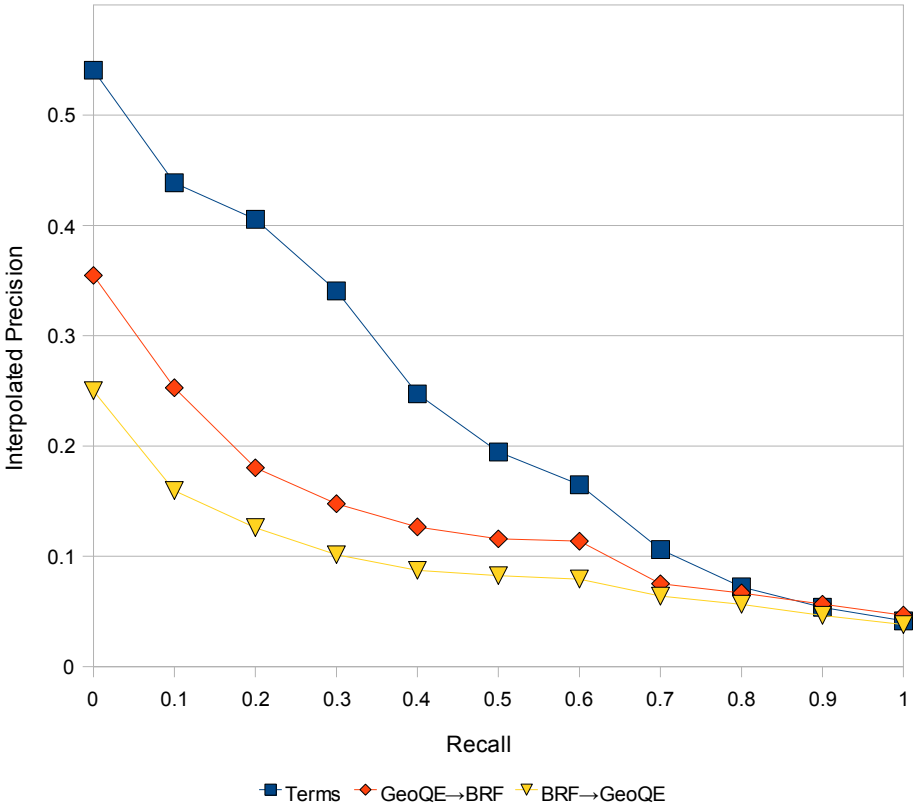


(a) Portuguese topics

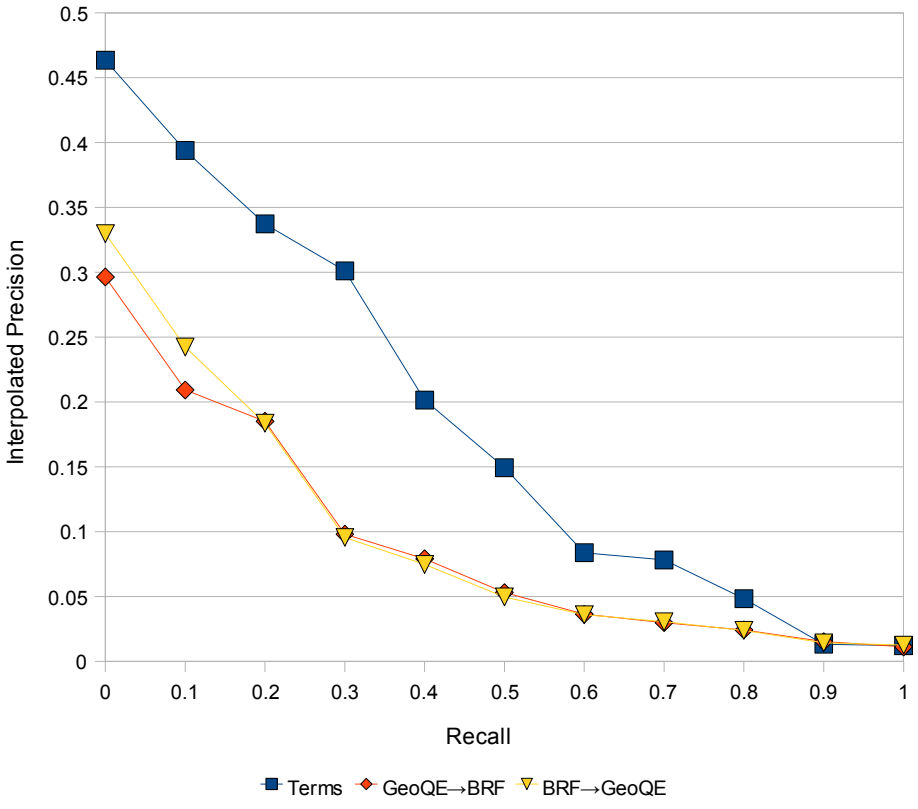


(b) English topics

Figure 6.4: MAP values of runs before query expansion

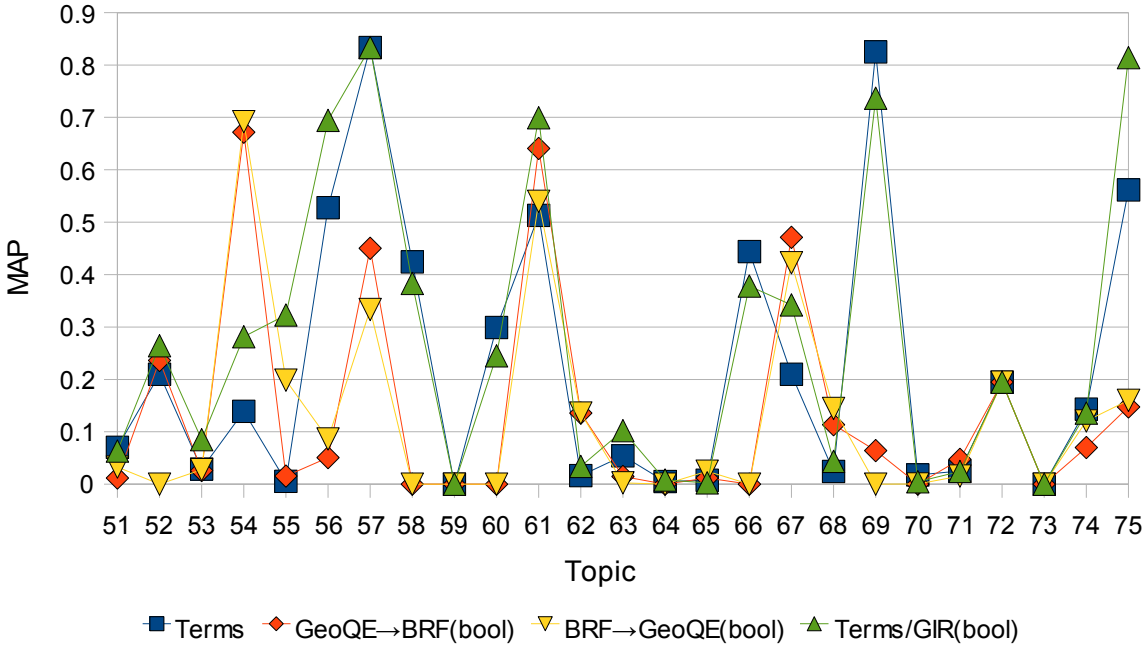


(a) Portuguese topics

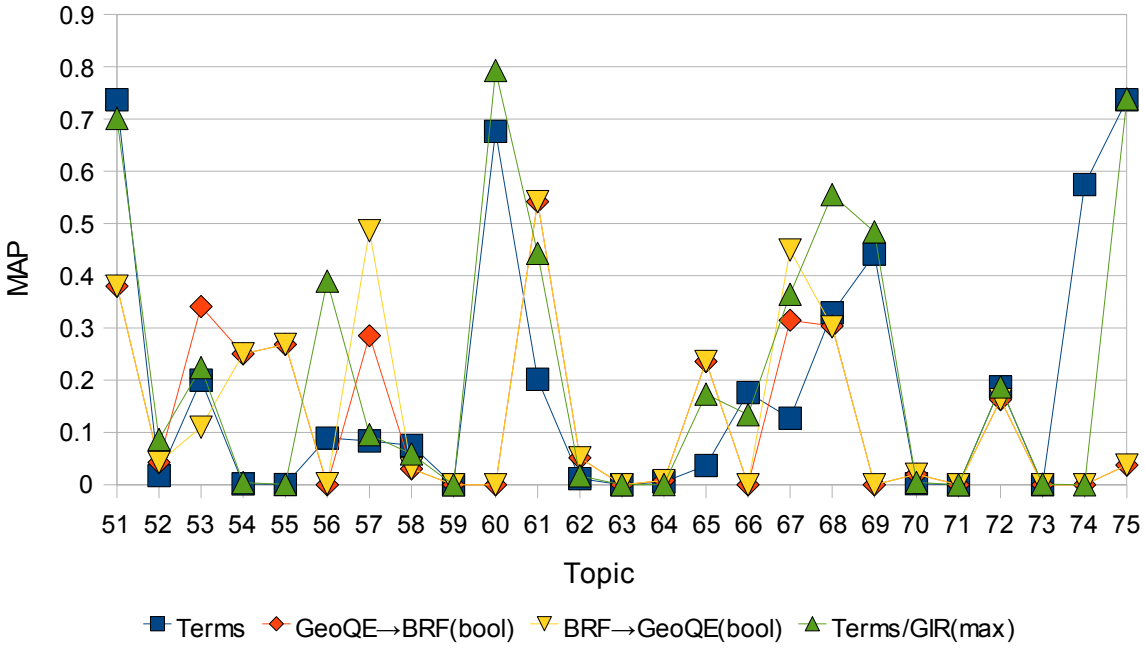


(b) English topics

Figure 6.5: Interpolated Precision-Recall curve for runs before query expansion



(a) Portuguese topics



(b) English topics

Figure 6.6: MAP values fo the best run for each strategy

Chapter 7

Conclusion

This work presented the new version of the indexing and ranking module of Tumba!, Sidra5, which implements a novel approach, based on geographic signatures, for the representation and reasoning of geographic information. Geographic signatures are the set of all geographic references, and their respective confidence measures, present on documents and queries alike. Their geographic references have been compared using several geographic similarity heuristics, which assign a value of likelihood between these two references.

Sidra5 was included in a prototype developed for participation in GeoCLEF 2007, to test the validity of the geographic signatures approach, evaluating if it obtained improved results relatively to a purely textual approach. The prototype also aimed at comparing several scoring metrics that have been proposed and comparing distinct query expansion strategies. GeoScore strategies compare a query and a document geographic signature and assign a score to the document that consists of geographic fitness relatively to the query.

The results of the query expansion strategies and those of the *Boolean* scoring seem to indicate that it is the geographic reasoning that improves the quality of the textual reasoning end, not the opposite. This underlines the importance of the textual component of queries on results, even in geographic environments. In either case, more studies have to be made to conclude about the real impact of this mutual contribution.

Geographic signatures are still in their infancy, yet the presented results and flexibility provided by the signatures show promises of improved geographic contextualization of search results. The potential of geographic signatures has just begun to be tackled and will continue to reveal itself as Sidra continues to mature and improve.

7.1 Future Work

This work has shown the validity of the geographic signature approach, yielding better search results than those provided by traditional IR approaches. Still, considering that this approach and the implemented systems are on their infancy, geographic signatures hold the promise of further improvements in search results.

The obtained results obtained gathered enough incentives to further mature the approach and proceed with the research and development on the following aspects:

Additional Geographic Scores: one inherent benefit of geographic signatures is the flexibility and diversity of distinct similarity evaluations possible. In this work, four geographic scoring methods were evaluated (Boolean, Max, Mean, Null), but more should be implemented and evaluated.

Query-dependent GeoScore Selection: not all geographic queries have the same geographic relationship and it is likely that some geographic scoring strategies obtain better search results in some geographic relations than others. The geographic scores could be evaluated relatively to each geographic relation so it will be possible to pick the appropriate GeoScore for each relation.

Optimal Weighting: weights used in both GeoScore and GeoSim have been guessed. Optimal GeoScore and GeoSim weights will have to be calculated. One possible way to obtain these scores could be to use a logistic regression, as proposed by Larson and Frontiera [2004].

New GeoSim Heuristics: the deceptioning results of $GeoScore_{mean}$, and $GeoScore_{bool}$ having better search results than $GeoScore_{max}$, could indicate an inadequacy of the geographic similarity heuristics. These heuristics were designed to a "one-scope per document" scenario, used in previous works, and do not seem to behave well with the "multiple-scope per document" used in the geographic signature approach.

Improved MG4J Use: MG4J is a full featured IR framework with extensive configurability. In the systems where Sidra5 was included, MG4J was not used to the full extent of its possibilities. As part of the ongoing effort to improve the quality and maturity of the implemented systems, some features of MG4J should be analyzed to allow reduced index size and improved query expansion support.

New Text Ranking Algorithms: the participation of the MG4J authors in the 2006 edition of TREC showed that the combination of text ranking algorithms can improve search results [Boldi and Vigna, 2006a]. Experiments could be

made to verify if the combination of Okapi BM25 and minimal-interval semantics ranking algorithms is also beneficial to the geographic signatures approach [Boldi and Vigna, 2006b].

Acknowledgements

I would like to thank my supervisor, Mário J. Silva, for his crucial guidance and his patience during all the dissertation preparation. I would like to thank Nuno Cardoso for the work on *QueOnde* and *QuerCol*, for the help in the result gathering and for the dissertation review; Marcírio Chaves for the work on *GKB* and for some crucial help; Joana Campos and Catarina Rodrigues for the work on *Faísca* and for managing the geographic data, respectively. I would also like to thank Daniel Gomes and Bruno Martins for their ideas and Leonardo Andrade for his guidance to understand previous Sidra versions. This work was supported by grant `POSI/SRI/40193/2001` (GREASE) from FCT, co-financed by POSI.

Appendix A

Data Tables & Topics

Table A.1: Distribution of D_{sig} by the number of geographic references

#	**Portuguese**	**English**
0	28804	18128
1	23538	15167
2	20229	13577
3	25317	14621
4	20031	13172
5	16300	10992
6	13644	9152
7	11126	7948
8	9049	7079
9	7253	5963
10	5837	5346
11	4973	4767
12	4101	4163
13	3324	3708
14	2619	3413
15	2238	3090
16	1812	2734
17	1583	2475
18	1282	2213
19	1073	2093
20	939	1856
21	787	1657
22	628	1464
23	518	1432
24	434	1238
25	378	1090
26	331	1052
27	315	893

Continued on next page

Table A.1 – continued from previous page

#	Portuguese	English
28	234	838
29	229	737
30	163	701
31	181	577
32	130	524
33	108	497
34	91	416
35	111	418
36	74	340
37	79	337
38	75	324
39	73	282
40	54	232
41	56	237
42	47	211
43	52	173
44	36	167
45	39	157
46	32	147
47	42	134
48	25	122
49	28	95
50	23	100
51	24	89
52	27	99
53	25	84
54	13	53
55	17	71
56	15	62
57	13	49
58	17	44
59	13	35
60	11	43
61	10	44
62	9	37
63	5	39
64	8	27
65	5	31
66	6	30
67	4	28
68	5	25
69	4	26

Continued on next page

Table A.1 – continued from previous page

#	Portuguese	English
70	8	20
71	2	20
72	4	9
73	4	14
74	2	6
75	3	12
76	2	9
77	5	13
78	2	11
79	0	15
80	3	9
81	1	6
82	1	8
83	0	5
84	2	5
85	1	7
86	3	7
87	0	5
88	0	5
89	0	4
90	1	5
91	3	1
92	1	7
93	0	5
94	0	4
95	1	3
96	0	2
97	0	5
98	0	1
99	1	6
100	0	5
101	0	2
102	0	2
103	1	0
104	0	5
105	0	3
106	0	3
107	1	4
108	1	5
109	0	2
110	0	1
111	0	0

Continued on next page

Table A.1 – continued from previous page

#	Portuguese	English
112	0	1
113	0	7
114	1	1
115	0	0
116	0	2
117	0	1
118	0	5
119	0	0
120	0	1
121	0	3
122	0	1
123	0	1
124	0	0
125	0	2
126	0	0
127	1	1
128	0	0
129	0	0
130	0	2
131	0	2
132	0	0
133	0	0
134	0	1
135	0	0
136	0	0
137	1	2
138	0	1
139	0	0
140	0	0
141	0	2
142	0	0
143	0	0
144	0	0
145	0	2
146	0	0
147	0	1
148	0	0
149	0	0
150	0	1
151	0	2
152	0	2
153	0	0

Continued on next page

Table A.1 – continued from previous page

#	Portuguese	English
154	0	1
155	0	0
156	0	0
157	0	0
158	0	0
159	0	1
160	0	0
161	0	0
162	0	0
163	0	0
164	0	0
165	0	0
166	0	0
167	0	0
168	0	0
169	0	0
170	0	0
171	0	0
172	0	0
173	0	1
174	0	0
175	0	0
176	0	0
177	0	0
178	0	1
179	0	0
180	0	0
181	0	1
182	0	0
183	0	0
184	0	0
185	0	0
186	0	0
187	0	0
188	0	0
189	1	0

Table A.2: Topics of the Web track from TREC 2004

Number	Topic
1	Electoral College
2	Ireland consular information sheet
3	Citizen attitudes toward prairie dogs
4	JPL stardust comet wild
5	American music
6	Philadelphia streets
7	Togo embassy
8	Philippines
9	Baltimore
10	well water contamination
11	Pileated woodpecker
12	oil petroleum resources
13	Eruption of Mount St. Helens
14	Club drugs
15	welfare reform
16	Sandhill cranes and the Platte river
17	Secure linux
18	Copyright basics
19	toxic waste
20	Tuskegee airmen observance
21	substance abuse
22	National atlas maps
23	Iraq Kuwait threat history
24	child care
25	History of Phoenix Symbol
26	The White House President Bush's cabinet
27	OPM New Retirees
28	FDA 2002 press releases
29	Grand canyon monitoring and research center
30	HIV/AIDS
31	origin of the universe
32	A history of American agriculture
33	teen pregnancy
34	NTP Herbal medicine factsheet
35	Religious Freedom amendment
36	the arts in education
37	magnetism
38	Ohio dams and locks
39	Child support enforcement cost/benefit
40	public school standards of performance
41	historic preservation

Continued on next page

Table A.2 – continued from previous page

Number	Topic
42	Sibir Air anatomy of a disaster
43	Vehicle registration California
44	Dash combination diet
45	faith-based initiatives
46	Local drinking water
47	medical residency
48	federal and state statistics
49	FCC consumer bureau
50	money laundering

Number	Topic
10.2452/51-GC	Extracção de petróleo e gás entre o Reino Unido e o continente europeu
10.2452/52-GC	Crime perto de Santo André
10.2452/53-GC	Investigação científica em universidades da costa leste da Escócia
10.2452/54-GC	Prejuízos causados por chuvas ácidas no Norte da Europa
10.2452/55-GC	Mortes causadas por avalanches na Europa excluindo os Alpes
10.2452/56-GC	Lagos com monstros
10.2452/57-GC	Uísque de ilhas escocesas
10.2452/58-GC	Problemas em aeroportos londrinos
10.2452/59-GC	Cidades em que houve reuniões da comunidade dos países andinos (CAN)
10.2452/60-GC	Baixas em Nagorno-Karabakh
10.2452/61-GC	Acidentes de avião perto de cidades russas
10.2452/62-GC	Reuniões da OSCE na Europa de Leste
10.2452/63-GC	Qualidade da água na costa mediterrânica
10.2452/64-GC	Acontecimentos desportivos na Suíça francesa
10.2452/65-GC	Eleições livres em África
10.2452/66-GC	Economia no Bósforo
10.2452/67-GC	Pistas em que Ayrton Senna correu em 1994
10.2452/68-GC	Rios com cheias
10.2452/69-GC	Morte nos Himalaias
10.2452/70-GC	Turismo no Norte da Itália
10.2452/71-GC	Problemas sociais na Grande Lisboa
10.2452/72-GC	Costas com tubarões
10.2452/73-GC	Ocorrências na catedral de São Paulo
10.2452/74-GC	Tráfego marítimo nas ilhas portuguesas
10.2452/75-GC	Violações dos direitos humanos na antiga Birmânia

Table A.3: Portuguese topics for GeoCLEF 2007

Number	Topic
10.2452/51-GC	Oil and gas extraction found between the UK and the European Continent
10.2452/52-GC	Crime near St Andrews
10.2452/53-GC	Scientific research at east coast Scottish Universities
10.2452/54-GC	Damage from acid rain in northern Europe
10.2452/55-GC	Deaths caused by avalanches occurring in Europe, but not in the Alps
10.2452/56-GC	Lakes with monsters
10.2452/57-GC	Whisky making in the Scottlsh Islands
10.2452/58-GC	Travel problems at major airports near to London
10.2452/59-GC	Meetings of the Andean Community of Nations (CAN)
10.2452/60-GC	Casualties in fights in Nagorno-Karabakh
10.2452/61-GC	Airplane crashes close to Russian cities
10.2452/62-GC	OSCE meetings in Eastern Europe
10.2452/63-GC	Water quality along coastlines of the Mediterranean Sea
10.2452/64-GC	Sport events in the french speaking part of Switzerland
10.2452/65-GC	Free elections in Africa
10.2452/66-GC	Economy at the Bosphorus
10.2452/67-GC	F1 circuits where Ayrton Senna competed in 1994
10.2452/68-GC	Rivers with floods
10.2452/69-GC	Death on the Himalaya
10.2452/70-GC	Tourist attractions in Northern Italy
10.2452/71-GC	Social problems in greater Lisbon
10.2452/72-GC	Beaches with sharks
10.2452/73-GC	Events at St. Paul's Cathedral
10.2452/74-GC	Ship traffic around the Portuguese islands
10.2452/75-GC	Violation of human rights in Burma

Table A.4: English topics for GeoCLEF 2007

Acronyms

APL Apache Public License

BRF Blind Relevance Feedback

BSD Berkeley Software Distribution

CLEF Cross Language Evaluation Forum

ConfMeas Confidence Measure

Docid Document identifier

D_{sig} Document geographic signature

GeoCLEF Geographic CLEF

Geo. QE Geographic Query Expansion

GeoScore Geographic Score

GeoSim Geographic Similarity

GIR Geographic Information Retrieval

GIS Geographic Information Systems

GKB Geographic Knowledge Base

IR Information Retrieval

JavaVM Java Virtual Machine

MAP Mean Average Precision

MG4J Managing Gigabytes for Java

MPL Mozilla Public License

P_r Precision

QE Query Expansion

Q_{sig} Query geographic signature

R_r Recall

TREC Text REtrieval Conference

Bibliography

Leonardo Andrade. Sidra 3: Sistema de Indexação no Motor de Busca Tumba! Technical report, Department of Informatics, University of Lisbon, July 2005.

Leonardo Andrade and Mário J. Silva. Indexing Structures for Geographic Web Retrieval. 2006a.

Leonardo Andrade and Mário J. Silva. Relevance Ranking for Geographic IR. 2006b.

P. Boldi and S. Vigna. MG4J at TREC 2006. 2006a.

P. Boldi and S. Vigna. Efficient Lazy Algorithms for Minimal-Interval Semantics. *Lecture Notes in Computer Science*, 4209:134, 2006b.

Sérgio Freitas Bruno Martins, Mário J. Silva and Ana P. Afonso. Handling Locations in Search Engine Queries. 2006.

Nuno Cardoso and Mário J. Silva. Query Expansion through Geographical Feature Types. 2007. To appear in GIR'07.

Nuno Cardoso, David Cruz, Marcírio Chaves, and Mário J. Silva. The University of Lisbon at GeoCLEF 2007. 2007. Unpublished.

M.B. Carmo, S. Freitas, A.P. Afonso, and A.P. Cláudio. Filtering Mechanisms for the Visualization of Geo-Referenced Information. *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 1–4, 2005.

M. Chaves, B. Martins, and M. J. Silva. GKB - Geographic Knowledge Base. DI/FCUL TR 05–12, Department of Informatics, University of Lisbon, July 2005. URL `http://www.di.fc.ul.pt/tech-reports/05-12.pdf`.

Marcirio Silveira Chaves, Catarina Rodrigues, and Mário J. Silva. Data Model for Geographic Ontologies Generation. In *XATA2007 - XML: Aplicações e Tecnologias Associadas*, pages 47–58, Lisbon, Portugal, February, 15-16 2007.

Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD '06: Proceedings of the 2006 ACM*

SIGMOD international conference on Management of data, pages 277–288, 2006. ISBN 1-59593-434-0.

Miguel Costa. Sidra: a flexible web search system. DI/FCUL TR 04–17, Department of Informatics, University of Lisbon, November 2004.

N. Craswell and D. Hawking. Overview of the TREC 2004 Web Track. *NIST Special Publication*, pages 500–261, 2004.

Otis Gospodnetić Erik Hatcher. *Lucene in Action*. Manning Publications, 2004.

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 Peters et al. [2007], pages 852–876. ISBN 978-3-540-74998-1.

Daniel Gomes. Tarântula – sistema de recolha de documentos da web. Technical report, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, August 2001. Report of the traineeship done by the author at the LaSIGE. In Portuguese.

Daniel Gomes and Mário J. Silva. Sistema de recolha de documentos da web (poster). In *CRC'01 - 4ª Conferência de Redes de Computadores*, November 2001.

Daniel Gomes, Joao P. Campos, and Mário J. Silva. Versus: a web repository. In *WDAS - Workshop on Distributed Data and Structures 2002*, Paris, France, March 2002.

Daniel Gomes, André L. Santos, and Mário J. Silva. Webstore: A manager for incremental storage of contents. DI/FCUL TR 04–15, Department of Informatics, University of Lisbon, November 2004. URL `http://www.di.fc.ul.pt/tech-reports/04-15.pdf`.

J. Graupmann and R. Schenkel. GeoSphereSearch: Context-Aware Geographic Web Search. *Proceedings of the 2006 Workshop on Geographic Information Retrieval, Seattle, USA*, pages 64–67, 2006.

D. Harman. The TREC conferences (pp. 9–28). *Proceedings of the Hypertext—Information Retrieval–Multimedia (HIM'95), Konstanz, Germany*, 1995.

G. Hobona, P. James, and D. Fairbairn. An Evaluation of a Multidimensional Visual Interface for Geographic Information Retrieval. *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 5–8, 2005.

Christian Sallaberry Julien Lesbegueries and Mauro Gaio. Associating Spatial Patterns to Text-units for Summarizing Geographic Information. 2006.

Ray R. Larson and Patricia Frontiera. Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In *Research and Advanced Technology for Digital Libraries*, volume 3232/2005 of *Lecture Notes in Computer Science*, pages 45–56. Springer Berlin, 2004.

D. Lin. An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In Alessandro Nardi and Carol Peters, editors, *Working Notes for the CLEF 2007 Workshop*, 19-21 de Setembro 2007.

A. Markowetz, Y.Y. Chen, T. Suel, X. Long, and B. Seeger. Design and Implementation of a Geographic Search Engine. *8th Int. Workshop on the Web and Databases (WebDB)*, 2005.

B. Martins, M.J. Silva, and L. Andrade. Indexing and Ranking in Geo-IR systems. *Proceedings of the 2005 workshop on Geographic information retrieval*, pages 31–34, 2005.

Bruno Martins, Nuno Cardoso, Marcírio Chaves, Leonardo Andrade, and Mário J. Silva. The University of Lisbon at GeoCLEF 2006. 2006.

Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors. *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*, volume 4730 of *Lecture Notes in Computer Science*, 2007. Springer. ISBN 978-3-540-74998-1.

Lee Wang Lihua Yue Qi Zhang, Xing Xie and Wei-Ying Ma. Detecting Geographical Serving Area of Web Resources. 2006.

SE Robertson, S. Walker, and GJF Jones. Hancock-Beaulieu, and Gatford, M. 1995. Okapi at TREC-3. *Proceedings of the Third Text Retrieval Conference (TREC-3)(Gaithersburg, Md.), NIST Special Publication*, pages 500–226, 1995.

J. J. Rocchio Jr. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

M. Sanderson and J. Kohler. Analyzing Geographic Queries. *SIGIR Workshop on Geographic Information Retrieval*, 2004.

M.J. Silva, B. Martins, M. Chaves, N. Cardoso, and A.P. Afonso. Adding Geographic Scopes to Web Resources. *CEUS-Computers, Environment and Urban Systems*, 30(378-399):93, 2006.

Ian Soboroff. Do trec web collections look like the web? *SIGIR Forum*, 36(2):23–31, 2002. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/792550.792554.

R. Song, JR Wen, S. Shi, G. Xin, TY Liu, T. Qin, X. Zheng, J. Zhang, G. Xue, and WY Ma. Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. *Proceedings of the Thirteenth Text REtrieval Conference Proceedings (TREC-2004)*, 2004.

Ligiane A. Souza, Clodoveu A. Davis Jr., Karla A. V. Borges, Tiago M. Delboni, and Alberto H. F. Laender. The role of gazetteers in geographic knowledge discovery on the web. *la-web*, 0:157–165, 2005. doi: http://doi.ieeecomputersociety.org/10.1109/LAWEB.2005.38.

T. Tezuka, T. Kurashima, and K. Tanaka. Toward Tighter Integration of Web Search with a Geographic Information System. *Proceedings of the 15th international conference on World Wide Web*, pages 277–286, 2006.

Arron R. Walker, Binh Pham, and Miles Moody. Spatial bayesian learning algorithms for geographic information retrieval. In *GIS '05: Proceedings of the 13th annual ACM international workshop on Geographic information systems*, pages 105–114, 2005. ISBN 1-59593-146-5.

J.M. Ware, I.D. Wilson, J.A. Ware, and C.B. Jones. A Tabu Search Approach to Automated Map Generalisation. *Proceedings of the tenth ACM international symposium on Advances in geographic information systems*, pages 101–106, 2002.

Nicola Stokes Yi Li, Alistair Moffat and Lawrence Cavedon. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. 2006.

V.W. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in Query Rewriting. *Proceedings of the 2006 Workshop on Geographic Information Retrieval, Seattle, USA*, pages 23–27, 2006.

Xing Xie Xufa Wang Wei-Ying Ma Zhisheng Li, Chong Wang. Indexing Implicit Locations for Geographic Information Retrieval. 2006.

Yinghua Zhou, Xing Xie, Chuang Wang, Yuchang Gong, and Wei-Ying Ma. Hybrid index structures for location-based web search. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 155–162, 2005. ISBN 1-59593-140-6.