Adding Geographic Scopes to Web Resources

Mário J. Silva Bruno Martins Marcirio Chaves
Ana Paula Afonso
Nuno Cardoso
Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal

Abstract

Many Web pages are rich in geographic information and primarily relevant to geographically limited communities. However, existing IR systems only recently began to offer local services and largely ignore geo-spatial information. This paper presents our work on automatically identifying the geographical scope of Web documents, which provides the means to develop retrieval tools that take the geographical context into consideration. Our approach makes extensive use of an ontology of geographical concepts, and includes a system architecture for extracting geographic information from large collections of Web documents. The proposed method involves recognising geographical references over the documents and assigning geographical scopes through a graph ranking algorithm. Initial evaluation results are encouraging, indicating the viability of this approach.

Keywords: Geographic Information Retrieval, Text and Web Mining

1 Introduction

Over the past decade, search engines evolved from using classic Information Retrieval (IR) models to inferring relevance from the analysis of the Web graph [Arasu et al., 2001]. We propose to further improve the quality of search systems, by integrating the geographical knowledge that can be inferred from Web resources. The problem of finding automatic ways to attach geographical scopes to these resources is indeed getting increasing attention [Naaman et al., 2004, Amitay et al., 2004, Jones et al., 2002, Ding et al., 2000], and IR systems that access resources on the basis of geographic context are starting to appear, both in the academic and commercial worlds (i.e. the SPIRIT project [Bucher et al., 2005, Vaid et al., 2005, Jones et al., 2002] or commercial systems such as `mirago.co.uk`, `local.yahoo.com` and `local.google.com`).

Unlike the spatial information used in a Geographic Information System (GIS), spatial information obtained from Web documents is often incomplete and fuzzy in nature. A GIS user can formulate data retrieval queries specifying complex spatial restrictions, while a search engine targets a wide variety of users who only provide simple queries. However, these users are also in need of retrieval mechanisms for queries with geo-spatial relationships. In order to support this, a first step concerns with assigning geographical scopes to Web resources, so that the same resources can latter be retrieved according to geographical criteria.

In this paper, we describe our research on the identification of geographic scopes for Web pages, defining scope as the region, if it exists, whose readers find the page more relevant than average. A geographic scope is specified as a relationship between an entity on the Web domain (a HTML page or a Website) and an entity in the geographic domain (such as a location or administrative region). The geographic scope of a Web entity has the same footprint as the associated geographic entity.

Once scopes are assigned, search engine queries may specify, explicitly or implicitly through context information, that the pages of interest must have a given geographic scope, or that the ranked retrieval process should assign more weight to those pages whose scope is closer to the searcher's geographic location. This can be useful for many tasks, as some Web resources are relevant primarily to communities within a specific geographic region. For instance, Web sites containing information on restaurants or theatres are primarily relevant to those living or visiting the neighbourhood.

Machine learning provides effective techniques for text classification, involving the automatic generation of classifiers from manually annotated training data [Yang, 1999]. However, with very few exceptions, most work in automated classification has ignored the presence of hierarchically structured classes and/or features. Typical methods treat the items to classify as a "bag-of-features," not taking into consideration the possible relationships that may exist among them. Classifying Web pages according to geographical scopes is also much harder than usual categorisation tasks. Much of the contextual information that could be used to disambiguate the geographical scopes in natural language texts is absent or external to the texts. The amount of training data per feature is also so low that there are no repeatable phenomena to base probabilistic methods on. For instance, the frequency of location names is in itself not sufficient for a good classification, as the same location name will usually not be repeated, even if the name is important. Because of the large number of classes and the huge number of relevant features needed to differentiate among them, we are also restricted to using very simple classifiers, both because of computational cost and to prevent complex models from overfitting.

We propose a novel approach for assigning geographical scopes to Web pages which has two major steps. The first concerns feature extraction, by recognising and disambiguating geographical references in the text. In the second step, we assign geo-scopes to the documents, using a graph-ranking algorithm similar to PageRank [Page et al., 1999] to combine and further disambiguate the available features. A central component of the whole process is an ontology, used as the source of names and relationships among geographical concepts.

We are studying the incorporation of geographic searches in the next version of *tumba!* (`www.tumba.pt`), a Web search engine [Silva, 2003], which indexes the sites of interest to the people related to Portugal [Gomes and Silva, 2005]. The statistics collected by our system over the past two years motivated this research:

- Geographic information is pervasive on the Web. An analysis of 3,775,611 pages found 8,147,120 references to the 308 Portuguese municipalities (administrative divisions of the territory corresponding to populated sub-regions of the Eurostat NUT3 areas), an average of 2.2 references per document [Martins and Silva, 2004b].

- Geographic entities are frequent in user queries. Approximately 4% of the queries logged by *tumba!* contain the names of the same 308 municipalities. If we considered names of localities, landmarks or streets, this percentage would increase.

It would certainly also increase if our engine considered geographic semantics and proximity when giving results. Other studies have also shown that geographically related queries represent a significant sub-set of the queries submitted to a global Web search engine [Sanderson and Kohler, 2004].

The rest of this paper is organised as follows: Section 2 lists the heuristics that can be used in a geographical retrieval system for the Web. Section 3 gives a general overview of our system. Section 4 describes how we find geographical references in the text and how we assign scopes to the documents. Section 5 presents the evaluation methodology and some initial experiments with our scope assignment method. Section 6 presents other projects addressing the same or similar problems. Finally, Section 7 presents our conclusions and outlines directions for future work.

2 Heuristics for Geo-Referencing Web Pages

The automated classification of Web pages according to their respective geographical scopes should consider all the different clues that are available. We can use different algorithms and techniques, including linguistic and statistical variants of natural language processing [Manning and Schütze, 1999], Web mining and analysis of Web graphs [Kosala and Blockeel, 2000, Chakrabarti et al., 1999], and Web information extraction [Grishman, 1997]. Our approach relies on a set of heuristics derived from observations of how people use geographical references in Web pages. We have three groups of heuristics, concerning explicit references to geographical concepts in the text, the specificities of the Web environment, and the usage of geographical references. They are described in detail in the rest of this Section.

2.1 Textual Information in Web Pages

- **If a geographical reference is present in a Web page, the scope of the page is related to the referenced region.** For instance having the sentence “city of Lisbon” in the text of a given Web page indicates that the scope of the page corresponds to the region of Lisbon. This is the major underlying assumption of our system. We assign geographic scopes to Web pages based on the geographical entities referenced in the pages.

- **The more often a term occurs in a document, the more likely it is important for that document** [Robertson and Jones, 1997]. Having the same geographic entity mentioned several times corresponds to a larger confidence in the document concerning that scope. For instance, if a document contains the word “Lisbon” twice and the word “Porto” only once, then it is more likely that the scope of the document corresponds to the region of Lisbon.

- **Text in a document is not all of the same importance [Cutler and Meng, 1997, Hill, 2000].** HTML defines a set of roles to which fragments of text can be assigned. Intuitively, terms appearing in different roles have a different importance (i.e. geographic references made in the title should be considered more important, and therefore weighted accordingly). For instance, if a document contains the word “Lisbon” in its title and the word “Porto” somewhere else on the page text, then it is more likely that the scope of the document corresponds to the region of Lisbon.

2.2 Hyperlinks and the Web Environment

- **The geographical scope of a document is related to the location of the Web server where it resides.** Although the location of hosting servers is somewhat unreliable, previous attempts indicate that using it for accessing the geographical scope of Web resources is a promising approach [Buyukokkten et al., 1999, Markowetz et al., 2005], at least when one considers large areas (i.e. countries). Some studies extended this view, using the `traceroute` utility to get location information for the Internet nodes connected to the server [Raz, 2004].

- **Geographical references extend their influence to more than the document they occur on.** Some references should spread to a wider extend than the Web page, in general the whole site. For instance, a common feature of company Web sites is a *contacts page*, which centralises all the possibilities for reaching an organisation, including postal addresses and telephone numbers. All site pages should, in principle, have the same geographical scope as the contacts page.

- **Hypertext links and geographical scopes are correlated.** This is derived from the topic locality assumption [Davison, 2000]. A link from document *A* to document *B* should mean that documents *A* and *B* are likely to have the same geographical scope, even if both documents are not part of the same Web site. We can propagate information from one page to those that are linked to/from it, in a way similar to the work presented in [Marchiori, 1998]. For instance, if several pages containing the word "Lisbon" link to a target page, then we can state that the scope of the target document also corresponds to the region of Lisbon, even if it does not contain any geographical terms in its text. The topic locality assumption only considers pages separated by a single link. More recent studies have shown that content relatedness decays with the distance from the start page [Menczer, 2002]. However, pages in a Web graph "further way" in terms of number of linkage "hops" tend to be less similar to the source document, and this should also apply to geographical scopes.

- **Geographical references in hyperlink anchor text are good indicators of the scope for the target page**. Text contained in hypertext anchors often contains good summaries for the target documents [Amitay, 1999]. This should also apply to geographic terms in link anchors. For instance, if a link to a given page contains the word "Lisbon" in its anchor text, then the scope of the target document should correspond to the region of Lisbon.

- **Some Web pages are more authoritative.** PageRank, for instance, measures page popularity through linkage information [Page et al., 1999]. The authority of a page is highly correlated with in-link counts [Kleinberg, 1999], and when assigning scopes, geographic references in pages with more in-links should weight more in the definition of the scope for the pages linked from them. For instance, if a very popular Web page links to another document with the word "Lisbon" in the anchor text, then we can state with a high degree of confidence that the scope of the target document corresponds to the region of Lisbon.

- **Linkage information should be explored in an aggregate sense.** A link from one page to another does not in itself provide a good indication that both pages refer to the same geographic scope, but, in an aggregate sense, if multiple pages of a given scope link to another page, that page should also have the same scope.

Moreover, links also give good indications on the “area of influence” of a page. Having many links from pages with scopes corresponding to a broad area of a country is different from having all the links from pages with the scope of a city in the same area. A similar assumption to this has already been used to find Web page’s popularity in a certain area [Inoue et al., 2002, Ding et al., 2000].

2.3 Usage of Geographical References

- **Each Web page has only one geographical scope.** This is similar to the “one sense per discourse” assumption, taken in many recognition and disambiguation systems [Gale et al., 1992]. It is possible in fact that one page concerns more than one scope (i.e. a page listing the hotels in a country could be seen as containing information about most of the cities in that country). However, with this assumption, we would use the whole country as the scope of that page.

- **If a document references a given location then, to some extent, it should also concern its sub-locations, adjacent locations, and related locations.** The different types of relationships between the geographical references made in the documents should be considered in disambiguating the available information. For instance, a page containing the word “Lisbon” should also be related to the a geographical scope corresponding to the “suburbs of Lisbon”.

- **Demographics data can be used to disambiguate geographical scopes.** Given two locations with the same name, the one with higher population is more frequently mentioned in Web pages. In the same way, when a geographic name in a text refers to several possible locations, the most likely scope of the page is the one of the most populated location. For instance, if a page contains the word “Lisbon” then it is more likely that it is referencing the capital of Portugal than another less populated location also named “Lisbon”.

- **In a place name hierarchy, names at higher levels are more likely to be referenced in Web pages and are also more recognisable by users.** As a result, hierarchy level information can be used to disambiguate geographical references. For instance, if a page contains the word “Lisbon” then it is more likely that it is referencing the capital of Portugal than a small village also named “Lisbon” in the interior of the country.

- **Geographical references identified in Web pages are more reliable than indirectly inferred geographical information.** Geographic names in a Web page or manually added meta-data are more important than those propagated through the linkage, or from data about the hosting Web site. For instance, if a page contains the word “Lisbon” then we can state that the scope of the page corresponds to the region of Lisbon with a higher confidence than if the page merely had an in-link from another page containing the word “Lisbon” in its text.

- **Multiple contexts that are in agreement provide increased confidence.** For instance, if a document explicitly mentions a place name, and the same place name can be associated with the hosting Web site, then it is probable that the page indeed concerns the scope corresponding to that place name.

3 System Overview

Our framework for mining Web data is shown in Figure 1. Building on the software architecture of our Web search engine (`www.tumba.pt`), the proposed framework relies on meta-data and XML standards, incorporating Semantic Web technologies such as Dublin Core and the Resource Description Framework (RDF) [Berners-Lee, 2000].

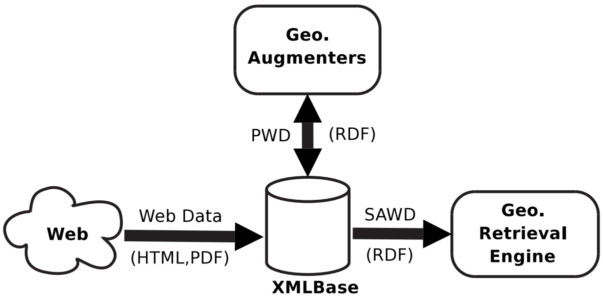

Figure 1: General Software Framework for Processing Web Resources.

Data is harvested into XMLBase, a Web data management system capable of processing large collections in parallel. Its components include a crawler, document processing tools [Martins and Silva, 2005b, Martins and Silva, 2005a], and separate data and meta-data repositories [Campos, 2003, Gomes et al., 2004]. While “crawling” Web documents (HTML, PDF, etc.), low-level data extraction (e.g. text tokens, n-grams, sentences, hypertext links, structural markup, and metatag information), and simple mining tasks (e.g. language identification) are also performed. This is an important first step, as handling Web data usually involves processing badly formatted information, with markup errors introduced by hand-editing documents or buggy authoring tools. The content of the Web documents is this way digested into RDF representations, which are then stored in the repository. We call these *Purified Web Documents* (PWD), to express that Web data, before becoming available for analysis, is cleaned into a collection of well-formed XML documents, organised under a common schema.

Purified Web Documents are the starting point of a chain of transformations that incorporate into the PWDs additional knowledge about the documents. Each of these transformers, called *augmenters* [Gruhl et al., 2004], can be thought as a domain-specific expert. Geographical knowledge is embedded within the PWDs as additional RDF resources, and we call these enriched documents *Scope Augmented Web Documents* (SAWD). Figure 2 details the process of assigning scopes to Web resources, implemented as two specific augmenters in the overall architecture presented above. One augmenter is responsible for recognising and disambiguating geographical references in the text, and the other is responsible for assign scopes using the geographical references and a graph-ranking algorithm similar to PageRank [Page et al., 1999].

Geographical domain knowledge is a key issue in the whole process. In our case, it is organised in a Geographic Knowledge Base (GKB), which integrates information from multiple public sources with geo-related and Web-related data. The purpose of the GKB is both to provide a common place for integrating data from the multiple external sources under a common schema, and supporting mechanisms for storing, maintaining, and exporting our knowledge on geographic concepts and Web resources. The exporting mechanism consists in generating OWL ontologies with the available

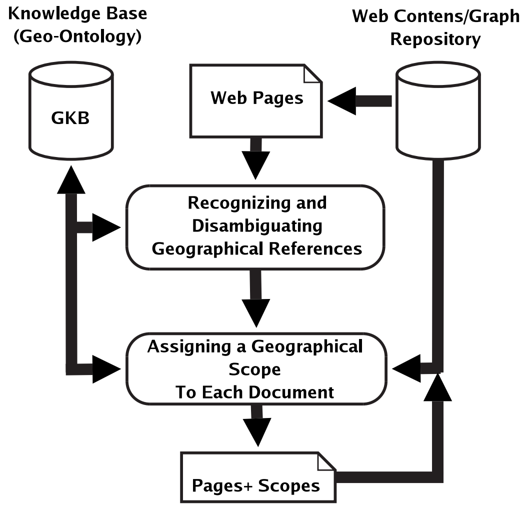

Figure 2: Assigning Geographic Scopes to Web Resources.

information, suitable to be used in the other software modules. The rest of this section details GKB and briefly describes a Web search interface that uses geographic scopes to retrieve information with geographic context.

3.1 GKB : The Geographic Knowledge Base

The geographic domain knowledge used throughout our system for assessing the geographic context of queries and documents is present in GKB, a repository of geographic data and knowledge rules relating the data [Chaves et al., 2005]. GKB includes all the information that is typically found in a gazetteer, such as names for places and other geographical features, type information (e.g. city, street, etc.), relationships between the geographic features, demographics data, and geographic codes such as postal codes and geographical coordinates. In addition to these data, GKB adds information about Internet domains and their relationship with the geographic features. For instance, GKB defines that the Internet domain `santiago-do-cacem.pt` corresponds to the Portuguese city of *Santiago do Cacém*.

GKB shares many resemblances with TGN, the Getty Thesaurus of Geographic Names (`http://www.getty.edu/research/conducting_research/vocabularies/tgn/`). TGN is a structured vocabulary including names and associated information about both current and historical places around the globe. However, GKB also adds Internet domain information and knowledge rules. In addition, we have created an instance with detailed data about Portugal, which is public and freely available as an OWL ontology `http://linguateca.di.fc.ul.pt/index.php?l=geonetpt`.

Figure 3 shows the common meta-model for storing the information in GKB. A *Feature* has a *Type*. The class *Name* contains all names identified for every feature. The classes *Relationship* and *Relationship_Type* capture the relations among features.

GKB manages not only geographic entities and relationships, but also the rules relating them. New knowledge is incorporated in GKB as rules, which may also be used by GKB programs to verify domain integrity rules and generate new relationships.

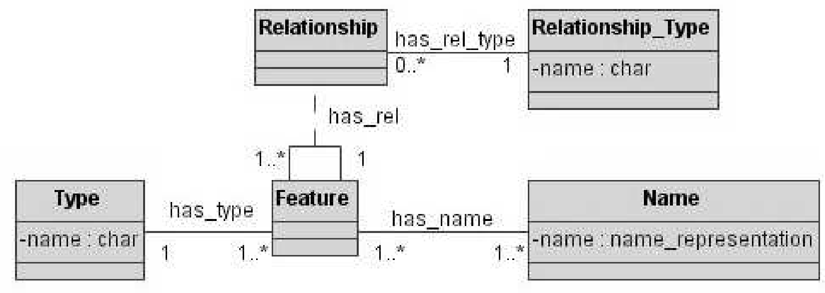

Figure 3: GKB information meta-model.

```
geoFeatureName(270,‘‘santiagodocacem’’).
geoFeatureName(270,‘‘santiagocacem’’).
geoFeatureName(270,‘‘santiago-do-cacem’’).
geoFeatureName(270,‘‘santiago-cacem’’).
geoFeatureType(270,‘‘CON’’).
netSiteSubDomain(33684,‘‘www’’).
netSitePrefix(33684,‘‘cm’’).
netSiteDomainToken(33684,‘‘santiago-do-cacem’’).
netSiteTLD(33684,‘‘pt’’).
```

Figure 4: ABox in DLs for the city of “Santiago do Cacém” (the values 270 and 33684 correspond to the feature identifiers in an instance of GKB holding these data)

To generate relationships, GKB receives the geographic data and rules in order to produce new relationships to be added to the relational database. For instance, we use Description Logics (DLs) [Baader et al., 2003] to specify rules and infer the scopes of DNS names. In general, the name given to a feature is represented in different ways. For multi-word names, such as *Santiago do Cacém*, the space character is the separator in texts. However, this character is invalid in URLs and domain names. A geographic name encoded in an URL has no spaces or may have hifens substituting for them or still may not have prepositions in its name.

Our rules system can automatically detect many sites with a well-defined geographic context, even when a geographic name is not represented in an obvious way. Figure 4 shows an extract of the world description (ABox) created in Description Logics from the data in GKB for the city of Santiago do Cacém. The world description has the different representations of geographic names. We represent the URL of web sites tokenized in three atomic concepts: sub-domain, domain and top-level domain (TLD). In addition, we also create the atomic concept `netSitePrefix`, which indicates the prefix to be used in a rule (in Portugal, most municipalities registered their domain as the city name prefixed by "cm-" or "mun-", the initials of *Câmara Municipal* or abbreviation of *Município*.

New knowledge is incorporated in GKB through rules, described in the Terminology Description (TBox in DLs): For instance, we express the knowledge that many Web sites of Portuguese `municipalities` are hosted in domains whose name contains the prefixes “cm-” or “mun-” by the following rule:

Site Type	# of sites	# of unifications
distritos	33	17 (52%)
municipalities	288	261 (90%)
freguesias	300	124 (41%)
basic schools	1955	124 (6%)
training centers	152	55 (36%)
high schools	402	105 (26%)

Table 1: Rule-based assigned scopes by GKB to sites of Portugal

Municipalities: $\texttt{hasScope(idN,idG)} \equiv \exists\texttt{netSiteDomainToken(idN,X)} \sqcap$
$$(\exists\texttt{netSitePrefix(idN,"cm")} \sqcup \\ \exists\texttt{netSitePrefix(idN,"mun"))} \sqcap \\ \exists\texttt{geoFeatureType(idG,"CON")} \sqcap \exists\texttt{geoFeatureName(idG,X)}.$$

The meaning of the previous rule is that there exists a `netSiteDomainToken X` which has the `netSitePrefix` “cm” or “mun” and corresponds to a `geoFeatureName X` of `geoFeatureType` “CON” (municipality). When an unification between the values `X` from `netSiteDomainToken` and `geoFeatureName` is found, we infer that the network feature represented by value `idN` has the geographic scope of the feature represented by the identifier `idG`.

We have been able to assign scopes to a large number of sites using the GKB instance of Portugal with this type of rules. Table 1 has some statistics. The number of sites identified for each feature type and the number of unifications obtained after the application of the rules are shown. For instance, Portugal has 308 municipalities and 288 of them have Web sites. For these, we assigned a geographic scope to 261 with the simple rule presented above. However, unifications do not always work, because occasionally the domain name of some of the sites does is not directly derived from the name of the corresponding geographic feature. For instance, the site `www.cm-ofrades.com` is about the municipality `Oliveira de Frades`.

3.2 The Geo-Tumba Geographical Web Search Engine

Assigning scopes to Web documents is not an end in itself. We are currently enhancing the tumba! Web search engine (`www.tumba.pt`) with more advanced retrieval tools that make use of the geographical scopes assigned to Web pages. For example, we can take into consideration the “geographical similarity” among the pages to cluster results related to the same scopes together, or we can rank the results according to the proximity with a given location.

The user interface of the new geographic search engine, called GeoTumba, is designed specifically for helping users in finding the most relevant Web pages that have a given spatial context. To meet this requirement, we introduced or modified three of the components of the typical Web search engine user interface: i) query formulation; ii) disambiguation of the spatial context of the query and iii) presentation of query results.

The query formulation interface of Geo-Tumba has two text input boxes instead of just one. The first enables the user to specify the query subject or concept, as before, and the second is reserved for input of the location information. Users create queries that are a combination of a location and particular items at those locations. To specify locations, users input textual descriptions such as a street addresses, a city or landmark name, a postal code, or latitude/longitude coordinates. The definition of the query location can also be given by clicking on an map.

The location information provided is used to infer the geographical context of the query. The specification of a geographical context requires the use of spatial relationships between the query subject and the query context. In our initial prototype, we only support the *in* spatial relationship (e.g., restaurants in Lisbon).

In disambiguating the geographical context of the query, the location information provided as input is initially checked for syntactical errors using a dictionary-based spelling checker [Martins and Silva, 2004a]. The dictionary contains the names of know locations on the geographic ontology and the corrected options are presented when location terms are not recognised. After this initial correction, we apply a named entity recognition (NER) procedure to recognise and disambiguate the geographical name in the query, again using the geographic ontology. If disambiguation is not possible, the user is prompted to disambiguate the location information provided by selecting one of the known contexts proposed.

Once the geographic scope of the query is resolved, it is submitted to the search engine and a list of the most relevant pages with scope matching the query is returned. To speed-up searches, we created two indexes: one keeps the pages associated to their corresponding scope, and the other associates scopes with their list of locations.

4 Assigning Scopes to Web Pages

As previously stated, the scope assignment process has two stages. First, we identify geographic names in the text of Web pages, weighting them according to occurrence frequency and HTML heuristics. In the second stage, we assign a final scope to each Web page, based on the geographical references found and on their relative weights. In the rest of this section, we describe both stages in detail.

4.1 Geographic NE Recognition and Disambiguation

The features for classification are extracted through a named entity recognition (NER) procedure particularly tailored to recognising and disambiguating geographical concepts over Web pages. Although NER is a familiar task within the Information Extrac-

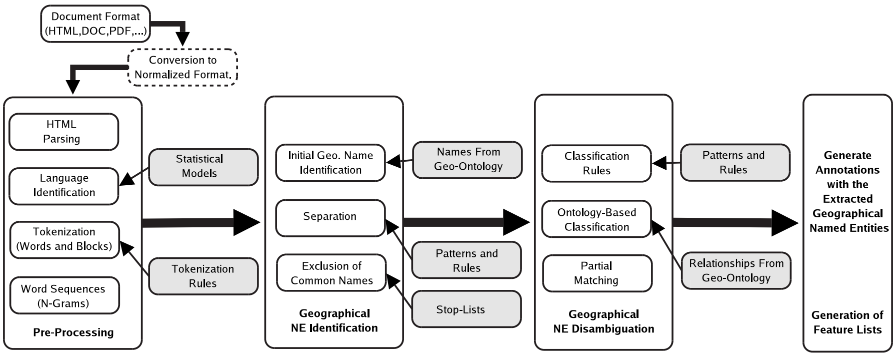
Document Format (HTML,DOC,PDF,...)
Conversion to Normalized Format.
HTML Parsing
Language Identification
Tokenization (Words and Blocks)
Word Sequences (N-Grams)
Pre-Processing
Statistical Models
Tokenization Rules
Initial Geo. Name Identification
Separation
Exclusion of Common Names
Geographical NE Identification
Names From Geo-Ontology
Patterns and Rules
Stop-Lists
Classification Rules
Ontology-Based Classification
Partial Matching
Geographical NE Disambiguation
Patterns and Rules
Relationships From Geo-Ontology
Generate Annotations with the Extracted Geographical Named Entities
Generation of Feature Lists

Figure 5: The geographical named entity recognition and disambiguation step.

tion community [Sang et al., 2003], our work advances the state of the art, as it presents a specific adaptation strategy to the domain of multilingual geographical references on the Web. Besides recognising place names, we try to normalise them in a way that specifically describes or even uniquely identifies the place in question, disambiguating them with respect to their specific type (e.g. city) and grounding them with features at the geographical ontology.

For each geographical reference, and besides the corresponding ontology features, we keep its associated occurrence frequency in the text, weighted according to HTML structure in a way similar to the scheme proposed in [Robertson et al., 2004]: structured HTML documents are first transformed into unstructured documents where different elements (i.e. the title) are repeated several times according to their importance. Hypertext anchors and other available meta-data information is also added to this more verbose document, which allows an uniform treatment of all geographical references. Although we left linkage out of our experiments, we could also add text from linking documents in the Web graph in this more verbose document. In Section 2 we listed some heuristics than can be used to account for geographical references from linked pages in the Web graph.

Figure 5 illustrates the geographical named entity recognition and disambiguation step, which follows the traditional architecture for NER systems by combining lexical resources with shallow processing operations. We highlighting its main conceptual stages, namely 1) pre-processing, 2) named entity identification, 3) named entity disambiguation, and 4) generation of feature lists.

Pre-processing starts by converting documents from multiple formats into HTML. It then parses documents to extract the textual information, using HTML markup to help in weighting text fragments and detecting text boundaries. A language guesser is used to classify the document's text [Martins and Silva, 2005b], and this information serves as the starting point for the more advanced processing operations. The identified textual segments are finally split into their constituent word *n*-grams, by moving a window over each text segment and taking all possible consecutive word sequences.

Named entity identification involves the detection of all possible *n*-grams that are likely to belong to a geographical reference. An initial identification applies language-specific patterns, which combine place names and context expressions with and without capitalisation (i.e. "city of Lisbon" or "Lisbon metropolitan area"). A separation sub-stage follows, in which *n*-grams that are likely to contain more than one named entity are detected and attachment problems are resolved. Finally, membership in exclusion lists is used to discard very frequent words that despite having a geographical connotation, are more frequently used in other contexts.

Named entity identification in itself does not derive the meaning of the expressions recognised. For instance, the expression "Campo Grande" refers both to a district in the city of Lisbon, and to a small town in another completely different region of Portugal. The named entity disambiguation stage addresses this issue, aiming to find the correct meaning for the expressions recognised. Classification rules use the same expressions that are used to recognise entities to solve simple cases (e.g. in "city of X", we know X is a city and not some other geographical feature). Ontology based classification uses the feature types and other contiguity measures to guess the correct type for a given reference (i.e. a one referent per discourse assumption so that place names throughout the same paragraph or the same line of an HTML table refer to the same or to geographically related locations). Finally, we compare slight word variations (i.e. one different character, one extra character or one less character) against references already disambiguated. In the cases not covered by the above heuristics, we keep the association of a reference to the several different possible items at the ontology. Some ambiguity problems can therefore persist at the end of the feature extraction phase.

4.2 Scope Assignment : Graph Ranking for Combining Features

Besides the ambiguity problems which may still persist after feature extraction, different geographical expressions (sometimes conflicting) can be associated with the same document. More then just counting the most frequent references, we need to combine the available information and further disambiguate among the different possible scope assignments that can be made for each document. This is the idea behind the second stage of our approach, which relies on the existence of a graph where the particular relationships between geographical concepts are specified.

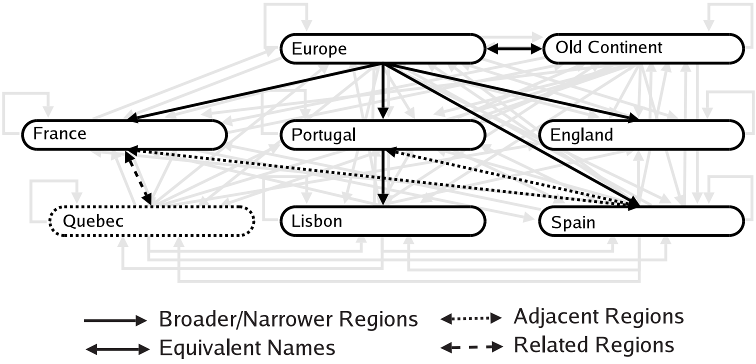

Figure 6: Generating a graph from the geographical ontology.

The geographical ontology provides the needed information – See Figure 6. We transform it to a suitable graph representation, weighting different semantic relationships (edges) according to their importance (i.e. equivalence relationships are more important than hierarchical relationships, which in turn are more important than adjacency relationships) and weighting different geographical concepts (nodes) according to the feature weights computed at the previous step. Importance scopes are then calculated for all the nodes in the graph. In the end, we select the highest ranked node as the scope for the document.

For the computation of "importance" scores, we use a variation of the popular PageRank ranking algorithm [Page et al., 1999]. PageRank determines the importance of a vertex using the collective knowledge expressed in the entire graph, recursively computing importance using the idea of "voting". The higher the number of "votes" (e.g. graph links) that are cast to a vertex, the higher its importance. Moreover, the importance of the vertex casting the vote determines the importance of the vote itself. There is a considerable amount of work focusing on all aspects of PageRank, namely stability, convergence speed, memory consumption, and the connectivity matrix properties [Khalil and Liu, 2004]. By using this formulation we can leverage on all these previous studies.

PageRank is traditionally computed through iterative solution methods. Formally, let $G = (V,E)$ be a directed graph with the set of nodes V and the set of edges E, where E is a subset of $V * V$. For a given node V_i, let $In(V_i) \subset V$ be the set of nodes that point to it, and let $Out(V_i) \subset V$ be the set of nodes that V_i points to. The values w_{ij} correspond to weights given to the edges connecting nodes V_i and V_j, and s_i correspond to weights given to each node V_i (the source strengths). Bellow we show the formula for graph-based ranking that takes into account edge and node weights when computing the score associated with a node in the graph. The ranking score of a node V_i is defined as:

$$S(V_i) = (1-d)s_i + d * \sum_{V_j \varepsilon In(V_i)} \frac{w_{ij}}{\sum_{v_k \varepsilon Out(V_j)} w_{jk}} S(V_j)$$

The parameter d is a damping factor that can be set between 0 and 1, integrating into the model the probability of jumping from a given node to another random node in the graph (e.g. having a document associated with a completely different feature than the ones we were able to extracted from it).

The source strengths s_i should be positive and satisfy the following condition:

$$|In(V_i)| = \sum_{j=1}^{|V|} s_i.$$

After a score is computed for each feature from the ontology, we select the most probable scope for the document, by taking the highest scoring feature or none if all are scored below a given threshold. The general procedure goes as follows:

1. Normalise the ranking scores obtained through the graph ranking algorithm.

2. If there are no features with a weight above the threshold, then no scope is selected.

3. From the set of features with the highest weight above the threshold:

   (a) If there is only one, return it as the scope for the document.

   (b) If there is more than one feature, but one of them corresponds to a generally broader concept in the ontology, return this broader feature as the scope.

   (c) If there is more than one feature, but they all have a common direct broader feature in the ontology, select this broader feature as the scope.

   (d) If there is more than one feature and no common direct broader concept exists, use demographics data to select the scope corresponding to the highest populated geographical region.

5 Evaluation Methodology and Initial Results

The subject of Geo-IR is still at an early stage of development, and limited evaluation has so far been performed on such systems. Advances in the area require an evaluation methodology, in order to measure and compare different techniques. A Geo-IR track at CLEF2005 was established as an initial experiment (see `http://ir.shef.ac.uk/geoclef2005/`). However, a complete Geo-IR system involves different components, which interdependently influence one-another and could benefit from a separate evaluation.

Disambiguating geographical references in the text and assigning documents with a corresponding geographical scope are two crucial steps in building a geographical retrieval tool. These two stages can be separately evaluated, and different settings and techniques could be used on each step. The performance of the extraction also dictates the performance of the scope assignments, giving an additional reason for a separate evaluation of each step.

Our approach interdependently relies on the used ontology, which influences the outcome of any experiment and should therefore be carefully analysed. Table 2 shows some statistics for two ontologies used in our experiments, namely a global one with names in different languages, and another specific for the Portuguese territory.

Portuguese ontology		Global Multilingual ontology	
Ontology Statistic	Count	Ontology Statistic	Count
Features	418065	Features	12293
Geographical Names	418460	Geographical Names	14305
Relationships	419072	Relationships	12258
Feature types	57	Feature types	7
Part-of relations	418340 (99.83%)	Part-of relations	12245 (99.89%)
Equivalence relations	395 (0.09%)	Equivalence relations	1814 (14.80%)
Adjacency relations	1132 (0.27%)	Adjacency relations	13 (0.10%)
NUT1	3	ISO-3166-1	239
NUT2	7	ISO-3166-2	3979
NUT3	30	Agglomerations	751
Districts	18	Places	3968
Islands	11	Admin. divisions	3111
Municipalities	308	Capital cities	233
Civil Parishes	3595	Continents	7
Zones	3594	Other	4
Localities	44386		
Street-like	146422		
Postal codes	219691		

Table 2: Statistics for the geographical ontologies.

The effectiveness of the feature extraction step was measured over newswire corpora used in previous NER evaluation efforts, and over a smaller collection of hand-annotated Web documents consisting of 20 HTML pages for each of 4 different languages. Table 5 summarises the obtained results, separating the simpler task of recognising the boundaries for geographical references in the text (recognition) from the tasks of distinguishing types and assigning the recognised references to the appropriate features at the ontology (disambiguation+grounding).

	Recognition			Disambiguation + Grounding		
Corpus	Precision	Recall	F_1	Precision	Recall	F_1
Portuguese text (HAREM)	0.89	0.68	0.77	-	-	-
English text (CoNLL-2003)	0.85	0.79	0.81	-	-	-
Spanish text (CoNLL-2002)	0.83	0.76	0.79	-	-	-
Portuguese HTML	0.90	0.76	0.82	0.89	0.76	0.81
English HTML	0.91	0.75	0.82	0.90	0.73	0.80
German HTML	0.79	0.72	0.91	0.77	0.70	0.73
Spanish HTML	0.86	0.75	0.80	0.83	0.72	0.77

Table 3: Evaluation results for the feature extraction stage.

The disambiguation and grounding tasks could not be evaluated over the newswire corpora, since they did not have locations tagged with their respective ontology feature. Although the Portuguese ontology included more place names, results in terms of recall are inferior to experiments using the smaller global ontology. Although further experiments are needed to confirm this, results indicate that recall does not improve considerably with the amount of available place names.

Performance over Web data is comparable to newswire texts, and our results are comparable to previous proposals. For instance, Nissim et al. experimented an off-the-shelf max-entropy tagger for recognising place names in Scottish historical documents [Nissim et al., 2004]. They achieved similar performances to state-of-the-art NER results (an f-score of 94.25%), but a preliminary experiment in recognizing specific types (i.e. cities) yielded a drop in performance of about 20%. Comparing our results with previous experiments in disambiguating place references – see Table 4 – can be a problem, as systems vary in the disambiguation performed and on the resources used for evaluation. For instance, some systems only classify references according to their correct type, while others also ground references to coordinates or to a gazetteer.

System	Classify	Ground	Evaluation Results
InfoXtract [Li et al., 2002]	✓	✓	94% accuracy
IDVL [Olligschlaeger and Hauptmann, 1999]	✓	✓	75% accuracy
Web-a-Where [Amitay et al., 2004]	✓	✓	63-82% accuracy
Smith and Mann [Smith and Mann, 2003]	✓		22-87% accuracy
Schilder et al. [Schilder et al., 2004]	✓	✓	74% f1-score
KIM system [Manov et al., 2003]	✓		88% f1-score
Nissim et al. [Nissim et al., 2004]	✓		f1-score around 75%
Leidner et al. [Leidner et al., 2003]	✓	✓	-
Metacarta [Rauch et al., 2003]	✓	✓	-

Table 4: Different systems handling geo-references in text.

As for the evaluation of the scope assignment process, the "gold-standard" collection consisted of 1000 pages from the ODP directory, located under the `Top:Regional` category. The specific branch of this category devoted to Portuguese pages with a coherent geographic scope was separately tested, together with some pages from Web sites for Portuguese municipalities (a total of 500 pages related to the Portuguese territory). Since the considered scopes are organised hierarchically, we evaluated the classification approach at different levels of "granularity." Our assumption was that assigning pages to a corresponding broader region should in principle be easier than assigning pages to a narrower area. Table 5 summarises the obtained results, with the last line showing the accuracy in the case of exact matches with the scope defined for the pages. The two separate columns for each dataset show the results for assigning scopes with basis on just the most frequently occurring feature (without the graph-based ranking algorithm to combine the available information), and with the application of the PageRank algorithm.

The relatively low accuracy confirms the difficulty inherent to the problem at hand. As expected, classification accuracy drops significantly if we consider narrower regions. These results can be confronted with those reported in other studies. Junyan et al. tried to classify pages according to three layers, namely nation, state and city. They used a hierarchical thesaurus of place names, achieving a best f-score of 86% [Ding et al., 2000]. Yamada et al. proposed a scheme to identify the geo-

Multilingual Ontology ODP Top:Regional			Portuguese Ontology ODP Top:Regional:Europe:Portugal		
	Measured Accuracy			Measured Accuracy	
Granularity	Most. Frequent	Ranking	Granularity	Most. Frequent	Ranking
Continent	91%	92%	NUT 1	84%	86%
Country	76%	85%	NUT 2	58%	65%
			NUT 3	44%	59%
			Municipalities	28%	31%
Exact Matches	67%	72%	Exact Matches	34%	53%

Table 5: Experimental results in different test scenarios.

graphical region mentioned in a Web page through a minimum bounding rectangle, reporting an accuracy of 88% [Yamada et al., 2002]. Amitay et al. presented an approach for finding the geographical focus of Web pages when several place names are mentioned in the text, using the immediate parent in a hierarchically structured gazetteer [Amitay et al., 2004]. ODP data was used for evaluation and the correct continent, country, city or exact scope were guessed respectively 96%, 93%, 32% and 38% of the times.

6 Related Work

The WebFountain project is an example of a computer cluster designed to analyze massive amounts of textual information, enabling the discovery of trends, patterns and relationships [Gruhl et al., 2004, Dill et al., 2003]. The architecture integrates applications that focus on specific tasks, using multi-disciplinary text mining approaches to extract data from Web resources. The Web data management architecture used in this work shares many ideas with WebFountain.

The SPIRIT project aims to develop a search engine aware of geographical terminology, using a geographical ontology [Fu et al., 2005, Jones et al., 2002]. Our framework differs on the emphasis put on geographic named entity recognition, the use of a graph ranking method for assigning a single scope to each document, the extensive use of names instead of coordinate information, and the availability of ontologies associating entities to geographic scopes.

Markowetz et al. describe an initial implementation of a geographical Web search engine for Germany [Markowetz et al., 2005]. Their system performs extraction of geographical features from Web documents, which are then mapped to coordinates and aggregated across link and site structure. The authors build on many previously proposed concepts and ideas. Examples are the work of Gravano and Shivakumar which introduced the notion of geo-scopes [Buyukokkten et al., 1999, Ding et al., 2000], or the Web-a-Where project which addresses geo-coding of Web documents for the entire globe [Amitay et al., 2004]. No empirical results are given, and again our framework differs on the use of a graph ranking method for assigning a single scope to each document, and on the extensive use of place names instead of spatial coordinates.

The NetGeo project also concerned geographic locations in the context of the Internet, collating information from multiple sources in order to assign the most probable longitude/latitude pair, together with the specificity and reliability of the location of IP addresses [Moore et al., 2000]. There is a correlation between the location of the server hosting a Web page and its scope. However, this is not the case in many situations. For instance, many Portuguese Websites are hosted in servers from U.S. providers or concentrated in locations with good network connectivity. Local sites from small populated cities are seldom hosted in that city. Another previous study successfully used the `traceroute` utility, to get location information for the Internet nodes connected to a Web server [Raz, 2004]. However, these projects aimed at finding coarse geographical scopes (such as states or countries), whereas our scopes are of a finer level of detail.

Previous studies have demonstrated that recognising geographical place names in text (usually called *geo-parsing*) is a crucial precondition for geo-referencing Web documents [Densham and Reid, 2003]. In language processing, the task of extracting and distinguishing different types of entities in text (i.e. names of people or organisations, dates and times, events, geographic features or even "non entities") is referred to as Named Entity Recognition (NER) [Sang et al., 2003, Palmer and Day, 1997]. The degree to which gazetteers help in identifying named entities seems to vary. For instance, Malouf found that gazetteers did not improve performance [Malouf, 2002], whereas others have gained significant improvements using gazetteers and simple "triggering" patterns [Carreras et al., 2002]. Mikheev et al. showed that a NER system could perform well even without gazetteers for most entity classes, although not for place names [Mikheev et al., 1999]. The same study by Mikheev et al. also showed that simple list lookup performs reasonably well for locations [Mikheev et al., 1999]. Ambiguity is the main problem associated with geographical references in text. The CoNLL shared task on named entity recognition concluded that ambiguity in geographical references is bi-directional, as the same name can be used for more than one location (referent ambiguity), and the same location can have more than one name (reference ambiguity) [Sang et al., 2003]. The former has another twist: the same name can be used for locations as well as for other class of entities, like persons or company names (referent class ambiguity).

More recently, the Workshop on the Analysis of Geographic References, held in conjunction with the 2003 NAACL-HLT conference, focused on topics concerning the recognition, disambiguation, normalisation, storage, and display of geographic references [Kornai and Sundheim, 2003]. The concerned problems are more complex than the simple recognition of place names in text, as NER in itself does not derive the meaning of the expressions recognised. To be useful, NER systems focusing on geographical concepts should handle the complex issues related to how people use geographical references. Place names lack precision in their meaning, and often vary with time, from person to person, and with the context in which they are used. Many times place names are simply temporary conventions, and people's vernacular geography if also often vague, as they may also be interested in the vicinity of a place without knowing its exact name. Not only spelling variations are common on geographical names, but also the places those names reference change in shape and size.

The GIPSY system for automatic geo-referencing of text uses a disambiguation method that incrementally constructs a polytope via merging flat polygons, in such a way that a third dimension is introduced for the intersecting area (polygon stacking). The authors report some initial experiments, but no exact evaluation figures are given [Woodruff and Plaunt, 1994]. Smith and Crane proposed an interesting resolution method, based on computing a centroid for all possible referents and discarding points that are more than two times the standard deviation away from the centroid [Smith and Crane, 2001]. They report f-scores between 0.81 and 0.96, but found that the centroid-based method lacks robustness. More recently, Yamada et al. proposed a scheme to identify the geographical region mentioned in a Web page through a minimum bounding rectangle, reporting an accuracy of 88% [Yamada et al., 2002]. Our work differs from these three previous proposals for disambiguating and combining geographical references in the sense that instead of using coordinates, it relies on semantic relationships between geographical concepts as provided by an ontology. The approach that is closest to ours is perhaps the work in the Web-a-Where project, which uses a hierarchical gazetteer [Amitay et al., 2004]

7 Conclusions

We presented our approach for automatically identifying geographic scopes in Web pages. A shared knowledge base is used to augment RDF-based descriptions of crawled Web pages with geographic meta-data. This work is part of a larger project which also involves the creation innovative IR algorithms in our Web search engine, using the notion of "geographical relatedness."

The assignment of geographical scopes to Web resources is made in two stages. In the first, geographical references occurring over the texts are recognised and disambiguated. In the second stage, a scope is assigned to the documents, combining the available information through a graph-ranking algorithm. A geographical ontology provides the names and the relationships between geographical concepts.

Both steps of the scope assignment process reflect a set of heuristics related to the different ways in which people place geographical context information on Web documents. Since many parameters are combined, a very important step concerns tuning the weights associated with the several parameters. For now, we are relaying on empirical tests and on published results from other experiments in the area of information retrieval. In the future, we plan on evaluating our approach using a systematic method to tune these parameters. Additional evaluation studies are in fact currently being performed (i.e. participation at GeoCLEF 2005), and there are also many ideas for future enhancements and for retrieval functionalities making use of geographical scopes.

8 Acknowledgements

Diana Santos, the coordinator of Linguateca, is a consultant to our project in language technology. Special thanks also to our colleagues and members of the tumba! development team Daniel Gomes, Nuno Seco, Leonardo Andrade and Sérgio Freitas, for their comments on early drafts of this paper.

This research was partially supported by the FCCN-Fundação para a Computação Científica Nacional, and FCT-Fundação para a Ciência e Tecnologia, under project grants POSI/SRI/40193/2001 (XMLBase) and POSI/SRI/47071/2002 (GREASE).

References

[Amitay, 1999] Amitay, E. (1999). *Words on the Web - Computer Mediated Communication*, chapter Anchors in context : A corpus analysis of web pages authoring conventions. Intellect Books.

[Amitay et al., 2004] Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging Web content. In *Proceedings of SIGIR-04, the 27th Conference on Research and Development in Information Retrieval*. ACM Press.

[Arasu et al., 2001] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology*, 1(1).

[Baader et al., 2003] Baader, F., Calvanese, D., Nardi, D., McGuinness, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook: Theory,Implementation and Applications*. Cambridge University Press.

[Berners-Lee, 2000] Berners-Lee, T. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper Business.

[Bucher et al., 2005] Bucher, B., Clough, P., Joho, H., Purves, R., and Syed, A. K. (2005). Geographic ir systems : Requirements and evaluation. In *Proceedings of ICC-05, the 12th International Cartographic Conference*.

[Buyukokkten et al., 1999] Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of Web pages. In *Proceedings of WebDB-99, the 1999 ACM SIGMOD Workshop on the Web and Databases*.

[Campos, 2003] Campos, J. P. (2003). Versus: A Web data repository with time support. DI/FCUL TR 03–08, Department of Informatics, University of Lisbon. Masters thesis.

[Carreras et al., 2002] Carreras, X., Marques, L., and Padro, L. (2002). Named entity extraction using AdaBoost. In *Proceedings of CoNLL-2002, the 6th Conference on Natural Language Learning*.

[Chakrabarti et al., 1999] Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., and Kleinberg, J. (1999). Mining the Web's link structure. *Computer*, 32(8).

[Chaves et al., 2005] Chaves, M., Silva, M. J., and Martins, B. (2005). A geographic knowledge base for semantic Web applications. In *Proceedings of SBBD-05, the 20th Brazilian Symposium on Databases*.

[Cutler and Meng, 1997] Cutler, Y. S. M. and Meng, W. (1997). Using the structure of HTML documents to improve retrieval. In *Proceedings of USITS-97, the 1st USENIX Symposium on Internet Technologies and Systems*.

[Davison, 2000] Davison, B. D. (2000). Topical locality in the Web. In *Proceedings of SIGIR-00, the 23rd Conference on Research and Development in Information Retrieval*.

[Densham and Reid, 2003] Densham, I. and Reid, J. (2003). A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the Workshop on The Analysis of Geographic References held at HTL/NAACL 2003*.

[Dill et al., 2003] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., and Zien, J. Y. (2003). SemTag and Seeker: Bootstrapping the semantic Web via automated semantic annotation. In *Proceedings of WWW-2003, the 12th World Wide Web Conference*.

[Ding et al., 2000] Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of Web resources. In *Proceedings of VLDB-00, the 26th Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc.

[Fu et al., 2005] Fu, G., Jones, C. B., and Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the web. In *Proceedings of DBA-2005, the 2005 IASTED International Conference on Databases and Applications*.

[Gale et al., 1992] Gale, W., Church, K., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*.

[Gomes et al., 2004] Gomes, D., Santos, A. L., and Silva, M. J. (2004). Webstore: A manager for incremental storage of contents. DI/FCUL TR 04–15, Department of Informatics, University of Lisbon.

[Gomes and Silva, 2005] Gomes, D. and Silva, M. J. (2005). Characterizing a national community web. *ACM Transactions on Internet Technology*, 5(3).

[Grishman, 1997] Grishman, R. (1997). Information extraction: Techniques and challenges. In Pazienza, M. T., editor, *Lecture Notes In Artificial Intelligence*, volume 1299. Springer-Verlag.

[Gruhl et al., 2004] Gruhl, D., Chavet, L., Gibson, D., Meyer, J., Pattanayak, P., Tomkins, A., and Zien, J. (2004). How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal - Utility Computing*, 43(1).

[Hill, 2000] Hill, C. (2000). Information space based on HTML structure. In Voorhees, E. M. and Harman, D. K., editors, *Proceedings of TREC-9, the 9th Text REtrieval Conference*. Department of Commerce of National Institute of Standards and Technology.

[Inoue et al., 2002] Inoue, Y., Lee, R., Takakura, H., and Kambayashi, Y. (2002). Web locality based ranking utilizing location names and link structure. In *Proceedings of W2GIS-2002, The Second International Workshop on Web and Wireless Geographical Information Systems, in conjunction with WISE-2002 3rd International Conference on Web Information Systems Engineering*.

[Jones et al., 2002] Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., and Weibel, R. (2002). Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project. In *Proceedings of SIGIR-02, the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press.

[Khalil and Liu, 2004] Khalil, A. and Liu, Y. (2004). Experiments with PageRank computation. Technical Report 603, Computer Science department at Indiana University.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5).

[Kornai and Sundheim, 2003] Kornai, A. and Sundheim, B., editors (2003). *Workshop on the Analysis of Geographic References*. (held in conjunction with NAACL-HLT 2003).

[Kosala and Blockeel, 2000] Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group (SIG) on Knowledge Discovery and Data Mining*, 2.

[Leidner et al., 2003] Leidner, J. L., Sinclair, G., and Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, Edmonton, Alberta, Canada.

[Li et al., 2002] Li, H., Srihari, K. R., Niu, C., and Li, W. (2002). Location normalization for information extraction. In *Proceedings of COLING-02, the 19th Conference on Computational Linguistics*.

[Malouf, 2002] Malouf, R. (2002). Markov models for language-independent named entity recognition. In *Proceedings of CoNLL-2002, the 6th Conference on Natural Language Learning*.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[Manov et al., 2003] Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., and Cunningham, H. (2003). Experiments with geographic knowledge for information extraction. In *Proceedings Workshop on Analysis of Geographic References*.

[Marchiori, 1998] Marchiori, M. (1998). The limits of Web metadata, and beyond. In *Proceedings of WWW-98, the 7th International World Wide Web Conference*.

[Markowetz et al., 2005] Markowetz, A., Chen, Y.-Y., Suel, T., Long, X., and Seeger, B. (2005). Design and implementation of a geographic search engine. Technical Report TR-CIS-2005-03, Department of Computer and Information Science of the Polytechnic University of Brooklyn.

[Martins and Silva, 2004a] Martins, B. and Silva, M. J. (2004a). Spelling correction for search engine queries. In *Proceedings of EsTAL-04, España for Natural Language Processing*.

[Martins and Silva, 2004b] Martins, B. and Silva, M. J. (2004b). A statistical study of the WPT-03 corpus. DI/FCUL TR 04–04, Department of Informatics, University of Lisbon.

[Martins and Silva, 2005a] Martins, B. and Silva, M. J. (2005a). Language identification in Web pages. In *Proceedings of ACM-SAC-DE-05, Document Engineering Track of the 20th Symposium on Applied Computing*.

[Martins and Silva, 2005b] Martins, B. and Silva, M. J. (2005b). WebCAT: A Web content analysis tool for ir applications. In *Proceedings of WI-2005, The 2005 IEEE/WIC/ACM International Conference on Web Intelligence*.

[Menczer, 2002] Menczer, F. (2002). Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*. Forthcoming.

[Mikheev et al., 1999] Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of EACL-99, the 9th Conference of the European Chapter of the Association for Computational Linguistics*.

[Moore et al., 2000] Moore, D., Periakaruppan, R., and Donohoe, J. (2000). Where in the world is netgeo.caida.org? In *Proceedings of INET-2000, the 10th Annual Internet Society Conference*.

[Naaman et al., 2004] Naaman, M., Song, Y. J., Paepcke, A., and Garcia-Molina, H. (2004). Automatic organization for digital photographs with geographic coordinates. In *Proceedings of JCDL-04, the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*.

[Nissim et al., 2004] Nissim, M., Matheson, C., and Reid, J. (2004). Recognising geographical entities in scottish historical documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

[Olligschlaeger and Hauptmann, 1999] Olligschlaeger, A. M. and Hauptmann, A. G. (1999). Multimodal information systems and GIS: The informedia digital video library. In *Proceedings of the 1999 ESRI User Conference*.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library. Working Paper.

[Palmer and Day, 1997] Palmer, D. D. and Day, D. S. (1997). A statistical profile of the named entity task. In *Proceedings of ANLP-97, the 5th conference on Applied Natural Language Processing*. Morgan Kaufmann Publishers Inc.

[Rauch et al., 2003] Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*.

[Raz, 2004] Raz, U. (2004). Finding a host's geographical location. Retrieved on 30th August 2004 from the World Wide Web: `http://www.private.org.il/IP2geo.html`.

[Robertson et al., 2004] Robertson, S., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of CIKM-04, the 13th Conference on Information and Knowledge Management*.

[Robertson and Jones, 1997] Robertson, S. E. and Jones, K. S. (1997). Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory.

[Sanderson and Kohler, 2004] Sanderson, M. and Kohler, J. (2004). Analyzing geographic queries. In *Proceedings of SIGIR-GIR-2004, the Workshop on Geographical IR held at the 27th Conference on Research and Development in Information Retrieval*.

[Sang et al., 2003] Sang, T. K., F., E., and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning*. Edmonton, Canada.

[Schilder et al., 2004] Schilder, F., Versley, Y., and Habel, C. (2004). Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

[Silva, 2003] Silva, M. J. (2003). The case for a portuguese Web search engine. In *Proceedings of ICWI-2003, the 2003 IADIS International Conference WWW/Internet*.

[Smith and Crane, 2001] Smith, D. A. and Crane, G. (2001). Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, London, UK. Springer-Verlag.

[Smith and Mann, 2003] Smith, D. A. and Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*.

[Vaid et al., 2005] Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In *Proceedings of SSTD-05, the 9th Symposium on Spatial and Temporal Databases*.

[Woodruff and Plaunt, 1994] Woodruff, A. and Plaunt, C. (1994). GIPSY: Automated geographic indexing of text documents. *Journal of the American Society of Information Science*, 45(9).

[Yamada et al., 2002] Yamada, N., Lee, R., Kambayashi, Y., and Takakura, H. (2002). Classification of web pages with geographic scope and level of details for mobile cache management. In *Proceedings of W2GIS-02, the 2nd Workshop on Web and Wireless Geographical Information Systems*.

[Yang, 1999] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1–2).