

**UTILIZANDO EVIDÊNCIA DA WIKIPEDIA
PARA RELACIONAR TEXTOS A LUGARES**

RAFAEL ODON DE ALENCAR

**UTILIZANDO EVIDÊNCIA DA WIKIPEDIA
PARA RELACIONAR TEXTOS A LUGARES**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: CLODOVEU AUGUSTO DAVIS JR.

Belo Horizonte
Setembro de 2011

RAFAEL ODON DE ALENCAR

**USING EVIDENCE FROM WIKIPEDIA TO
RELATE TEXT TO PLACES**

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: CLODOVEU AUGUSTO DAVIS JR.

Belo Horizonte
September 2011

© 2011, Rafael Odon de Alencar.
Todos os direitos reservados.

Alencar, Rafael Odon de

A368u Utilizando Evidência da Wikipedia para Relacionar
Textos a Lugares / Rafael Odon de Alencar. — Belo
Horizonte, 2011.
xxiv, 54 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de
Minas Gerais. Departamento de Ciência da
Computação.

Orientador: Clodoveu Augusto Davis Jr.

1. Computação - Teses. 2. Sistemas de informação
geográfica - Teses. 3. Sistemas de recuperação da
informação - Teses. I.Orientador. II. Título.

CDU 519.6*72 (043)



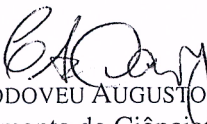
UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

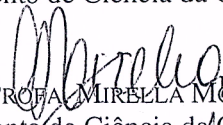
FOLHA DE APROVAÇÃO

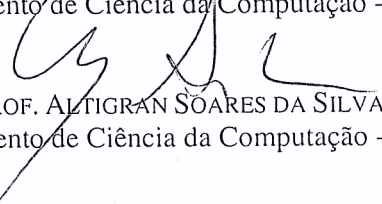
Utilizando Evidência da wikipedia para relacionar textos a lugares

RAFAEL ODON DE ALENCAR

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:


PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG


PROF. MIRELLA MOURA MORO
Departamento de Ciência da Computação - UFMG


PROF. ALTIGRAN SOARES DA SILVA
Departamento de Ciência da Computação - UFAM

Belo Horizonte, 29 de julho de 2011.

I dedicate this work to everyone who deserves a high quality study, but unfairly has no chance to get it. In their honor, I hope to have fully accomplished it with my commitments.

Acknowledgments

First of all, I thank God who assured me health and peace so that I could fight for this project! I also thank my advisor Clodoveu Augusto Davis Jr for being such a great professional and person, demonstrating wisdom and patience especially in the last months. I thank my mates from the Database Laboratory for the living together, for the exchange of ideas, and for all the laughs and friendship which made it all worthwhile. I also thank my family for their affection and specially for the education and values that I've inherited from them. I thank my girlfriend Carolina for her continued support, trust, understanding and complicity!

*“If you don’t make mistakes, you’re not working on hard enough problems.
And that’s a big mistake.”*

(Frank Wilczek)

Resumo

Para obter resultados relacionados a lugares em sistemas de buscas, usuários geralmente incluem termos geográficos nas consultas. Trabalhos anteriores mostram que consultas com intenção geográfica correspondem a uma parcela significativa da demanda submetida a máquinas de busca. Dessa forma, é importante abordar o desafio da associação automática entre lugares e documentos a fim de responder adequadamente a essas consultas. Esta dissertação descreve estratégias automáticas para a determinação do escopo geográfico de textos usando a Wikipedia como uma fonte de referências geográficas diretas e indiretas.

Em primeiro lugar, é proposto um método que utiliza evidências textuais da Wikipedia para classificar textos em lugares. São utilizados títulos de artigos e conexões entre os eles para estabelecer uma rede semântica que fornece informação para a classificação. Experimentos com notícias classificadas nos estados brasileiros demonstrando o potencial e as limitações da técnica. Outra proposta descreve uma estratégia para a marcação (*tagging*) de textos com múltiplos nomes de lugares. Dessa vez, utiliza-se uma técnica de identificação de tópicos (*topic indexing*) que considera os artigos da Wikipedia como um vocabulário controlado. Ao identificar esses tópicos no texto, conecta-se ele à rede semântica de artigos da Wikipedia, o que permite realizar operações no grafo e encontrar lugares relacionados ao tema tratado. É apresentada uma avaliação experimental com documentos previamente marcados, demonstrando a viabilidade da proposta e abrindo caminho para outras pesquisas em torno de *geotagging* baseado em redes semânticas.

Os resultados apresentados demonstram a viabilidade do uso da Wikipedia como uma fonte alternativa de referências geográficas. A principal vantagem do método proposto é o uso das informações livres, atualizadas e amplas da enciclopédia digital. Finalmente, considera-se que a introdução da Wikipedia na análise de elementos geográficos de textos pode ser encarada tanto como uma alternativa quanto como uma extensão ao uso dos dicionários toponímicos (*gazetteers*) em tarefas de recuperação de informação geográfica.

Abstract

Obtaining or approximating a geographic location for search results often motivates users to include place names and other geography-related terms in their queries. Previous work shows that queries that include geography-related terms correspond to a significant share of the users' demand. Therefore, it is important to recognize the association of documents to places in order to adequately respond to such queries. This dissertation describes strategies for the geographic scope computation, using Wikipedia as an alternative source of direct and indirect geographic references.

First we propose to perform a text classification task on geography-related classes, using textual evidence extracted from Wikipedia. We use terms that correspond to articles titles and the connections between articles in Wikipedia's graph to establish a semantic network from which classification features are generated. Results of experiments using a news data-set, classified over Brazilian states, show that such terms constitute a valid evidence set for the geographic classification of documents, and demonstrate the potential of this technique for text classification. Another proposal describes a strategy for tagging documents with multiple place names, according to the geographic context of their textual content, using a topic indexing technique that considers Wikipedia articles as a controlled vocabulary. By identifying those topics in the text, we connect documents with the Wikipedia semantic network of articles, allowing us to perform operations on Wikipedia's graph and find related places. We present an experimental evaluation on documents tagged as Brazilian states, demonstrating the feasibility of our proposal and opening the way to further research on geotagging based on semantic networks.

Our results demonstrate the feasibility of using Wikipedia as an alternative source of geographical references. The method's main advantage is the use of free, up-to-date and wide knowledge and information from the digital encyclopedia. Finally, the Wikipedia introduction to the geographic text analysis can be faced as both an alternative and a extension to using geographical dictionaries (i. e. gazetteers).

List of Figures

1.1	Textual content example – a news article about victims of the September 11 terrorist attacks in New York City that fails to mention the city’s name	3
3.1	Example of Wikipedia geographic categories	16
3.2	Example of an infobox from an Wikipedia article on Belo Horizonte city	17
3.3	Wikipedia graph example	22
3.4	Term weights in the inlinks and outlinks sets for each place	22
3.5	Sample documents for classification	23
3.6	Discovering places related to a wikified text in Wikipedia’s graph.	29
4.1	Accuracy versus training set sizes	37
4.2	Average accuracy versus number of classes	38
4.3	Geotagger performance for different exponential decay values for a <i>maxdepth</i> of 2	42

List of Tables

1.1	Direct and indirect geographic references found on September 11 news text (Figure 1.1).	4
3.1	Features in the first strategy, and calculated classes for each document. . .	23
3.2	Features in the second strategy and classification results	24
4.1	Details about the test collection built with local news from each one of the 27 states from Brazil.	44
4.2	Accuracy for different training set sizes classifying the same test set	45
4.3	TF-IDF strategy average accuracy considering a varying number of classes	45
4.4	Wikipedia strategy average accuracy considering a varying number of classes	45
4.5	Acceptance and accuracy for different <i>maxdepth</i> levels.	45
4.6	Geotagger performance for different exponential decay values for a <i>maxdepth</i> of 2	46

Contents

Acknowledgments	xi
Resumo	xv
Abstract	xvii
List of Figures	xix
List of Tables	xxi
1 Introduction	1
1.1 Motivation	1
1.2 Hypothesis	5
1.3 Objectives	5
1.4 Text Organization	5
2 Related Work	7
2.1 Geographic Information Retrieval	7
2.2 Geographic Context in Documents	8
2.3 Wikipedia as a Source of Information and Knowledge	10
3 Using Evidence from Wikipedia to Automatically Relate Texts to Places	13
3.1 Wikipedia as Source of Geographic Evidence	14
3.2 Strategy 1: Identifying Geographic Context through Classifiers with Wikipedia	18
3.3 Strategy 2: Identifying Geographic Context through Topic Indexing with Wikipedia	24
3.3.1 Defining a List of Target Places within Wikipedia	26
3.3.2 Topic Indexing with Wikipedia	26

3.3.3	Navigating in Wikipedia’s Graph to Discover Places	28
3.3.4	Evaluating Certainty and Choosing Responses	30
4	Experimental Evaluation	33
4.1	Building the Test Collection	33
4.2	Results for the Classifier Strategy	34
4.2.1	Baseline: Traditional Automatic Text Classification	35
4.2.2	Experiments and Results	36
4.3	Results for the Topic Indexing Strategy	40
4.3.1	Search Depth Influence on Performance	40
4.3.2	Manual Evaluation	42
5	Conclusions and Future Work	47
	References	51

Chapter 1

Introduction

1.1 Motivation

The last few decades brought us to an interesting time for research on data and information management. Computer applications have been developed to help us with several activities. Meanwhile, mass storage devices got cheap and large, allowing those applications to ignore how much data they produce. Computer networks also have their share, especially the World Wide Web, since they make it possible to millions of users to act as producers and consumers of such information. These and other factors motivate advances on information retrieval. Keyword-based search engines such as Google¹, Yahoo!² and Bing³ are widely used, and represent a billion-dollar business. But many challenges still exist. Taking into consideration the geographic meaning of keywords is one of them.

According to Wang et al. (2005), most human activities happen in a localized perspective, and that can also be reflected on the Internet. Previous works demonstrate that a large share of user queries on search engines have some geographic intention (Sanderson and Han, 2007; Delboni et al., 2007). Users often include place names and other geography-related terms in their queries, and keyword-based search engines could improve their results by considering the spatial dimension besides the classic keyword approach for query and document similarity (Baeza-Yates and Ribeiro-Neto, 1999). In order to deal with this challenge, we need to rely on some formal geographic annotation on web resources. However, we cannot expect that every retrievable resource will contain such metadata. This suggests the need for an automatic way to identify

¹<http://www.google.com>

²<http://www.yahoo.com>

³<http://www.bing.com>

the relation of Web pages to places.

Wang et al. (2005) consider three different ways to determine the geographic location(s) associated with a web document. First, it is possible to obtain an approximate location of the Web server, as informed by services that relate an IP address to a pair of coordinates, such as GeoIP⁴. Second, the location can be inferred by looking at concentrations of users that access the document and their IP-determined locations, or by the location of documents that refer to it. Third, the location can be inferred by analyzing the textual content of the document. All three aspects have their importance, but we are particularly interested in the last one. Notice that the geographic scope of the textual content can be totally different from the place where the document is served. Also, the location of visitors could be biased, since content providers can be more popular in specific places. For instance, imagine a news website from a Brazilian news agency located in United States of America, hosting its website in a European web server, and posting news about a fact in China. Much of the geography present in this complex example is not relevant for those who are only interested in Chinese facts, and we should consider that, in a globalized perspective, relevant information can come from anywhere. Since most of the information on the Internet is written down as natural language text, we also consider that textual analysis is a rich field of possibilities.

Much recent work follows this direction, with subjects such as the identification of geographic context in Web documents or the association of place names to Web pages (Jones et al., 2008; Schockaert et al., 2008; Silva et al., 2006) or simply *geotagging/geoparsing* (Amitay et al., 2004; Guillén, 2008). Successfully accomplishing this task enables us to associate places to online content, giving us means to enhance current indexing and retrieval mechanisms, so that people can search for textual content that fall within a delimited geographic scope (i.e. perform local search (Himmelstein, 2005; Schockaert et al., 2008)), or even filter content based on regional interests. Service providers would be able to perform geographically-focused advertising and to develop novel ranking strategies for search engines.

In practice, to geoparse content, three basic activities must be performed: identify place names candidates in the text; disambiguate them, since they can refer to multiple places or even to non-geographic entities; and finally assign a localization to the references, such as latitude and longitude coordinates (Amitay et al., 2004; Zong et al., 2005). In this case, gazetteers play the role of geographical dictionaries used as a source of official and alternative place names, mapped to a geographic infrastructure

⁴<http://www.maxmind.com/app/ip-location>

(i.e. points, lines and polygons described with coordinates) (Machado et al., 2010). This place name recognition approach can be very useful to find direct references in texts, but sometimes we must consider also the presence of indirect references.

Indirect references can lead us to infer the relation of the text with unmentioned places. Figure 1.1 is a news related to victims of the September 11 terrorist attack in New York. Curiously, the city name “New York” is never mentioned in the text. Although it seems quite easy for human readers to find out the text relation to New York, this task can be not so simple for computers.

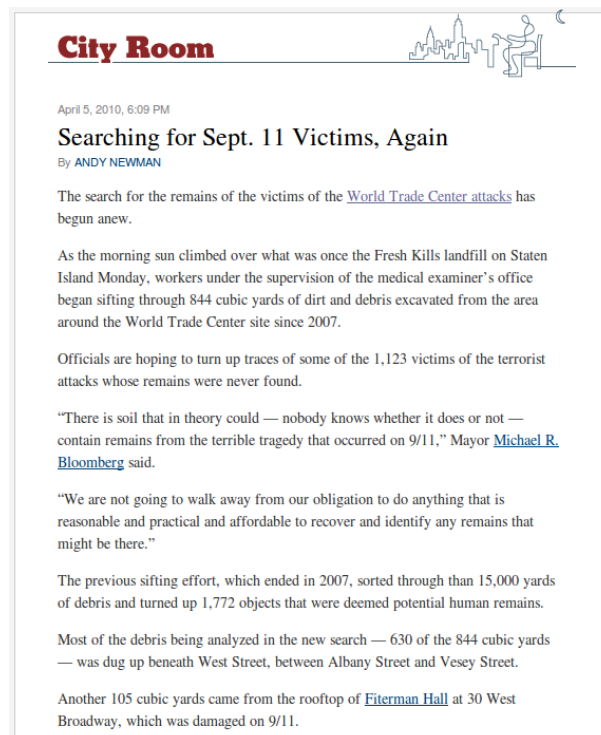


Figure 1.1. Textual content example – a news article about victims of the September 11 terrorist attacks in New York City that fails to mention the city's name

We can perform a quick manual analysis of the news content in Figure 1.1, and try to recognize geographic references. Some of them are listed in Table 1.1. Notice that the first references, such as the name of neighborhoods, streets and important buildings, could be automatically recognized with the help of a proper gazetteer containing intra-urban data about New York City. Nevertheless, the last ones, such as the current mayor's name, important dates in different formats and a relevant topic to the city's history (*Terrorist Attack*), are harder to deal with. Some of the difficulties are: (1) to propose a gazetteer with this level of detail for a considerable number of places, and (2) to keep such gazetteer up-to-date. In this work, we propose that this kind of non-

#	Reference	Description
1	Staten Island	Neighborhood
2	West Street	Street
3	Albany Street	Street
4	Vesey Street	Street
5	West Broadway	Street
6	World Trade Center	Building
7	Fiterman Hall	Building
8	Fresh Kills Landfill	Landfill
9	Michael R. Bloomberg	Mayor
10	Sept. 11	Date
11	9/11	Date
12	Terrorist Attacks	Topic

Table 1.1. Direct and indirect geographic references found on September 11 news text (Figure 1.1).

trivial geographic references could be fetched from third-party Knowledge Bases (KB). By exploring the KB's semantic capabilities we can try to make computers understand the relation among places and several types of topics. Since it is also important to keep information up-to-date, this KB should also be constantly enriched and evaluated.

Recent works have shown that Wikipedia⁵ can play the role of a Knowledge Base (Kasneci et al., 2009). Wikipedia is an online encyclopedia to which any visitor can contribute. Its pages represent concepts, and the hyperlinks among them allow users to navigate and explore a rich network of topics. Pages can be organized in categories, helping to group and list concepts in common branches of knowledge. Important structures present in the pages' source code can be used to gather domain specific information (e.g. retrieve the current mayor's name in cities pages). Also, the text used to explain the concepts in each page is a rich source of expressions and terms to feed both natural language processing tasks and probabilistic models for automatic text analysis. Such features suggests that Wikipedia can be used, by itself or along with official gazetteers, to improve the identification of direct and indirect geographic references in text. Wikipedia is also a good source for geographic knowledge because it is available in multiple languages, thus allowing the development of language-independent techniques. It is constructed and maintained by a large community interested in keeping information up-to-date. Therefore it is a suitable candidate for supplying semantic knowledge to information retrieval.

⁵<http://www.wikipedia.org>

1.2 Hypothesis

In this work, we explore the hypothesis that it is possible to use Wikipedia as an alternative source of geographic evidence to help dealing with the problem of automatically associating places to texts.

1.3 Objectives

In order to verify the hypothesis, we pursue two ideas in this dissertation. First, we formulate the location recognition problem as a classification problem, and propose a solution based on machine learning which uses Wikipedia contents to generate features that describe the presence of keywords associated to places in texts. Second, we use a topic indexing technique to discover relevant Wikipedia pages as topics in text.

In the first approach, a classifier must decide whether a text is more likely related to a certain place than to others, based on supplied training data, thus choosing a predominant geographic scope. The second approach intends to achieve a direct connection between terms and expressions found in text and the Wikipedia's network of topics. We then investigate if this network can provide enough information to build a list of directly and indirectly related places.

1.4 Text Organization

This dissertation is organized as follows. Chapter 2 lists related work. Chapter 3 describes our proposals to use Wikipedia as an alternative source for the identification of direct and indirect geographic references in text. Chapter 4 exposes a methodology to evaluate the feasibility and precision of our proposals, while Chapter 5 presents conclusions and future work.

Chapter 2

Related Work

2.1 Geographic Information Retrieval

The field of Information Retrieval (IR) comprehends problems on how to store and access information items, in such a way that users can query this information. Given the user needs, expressed through queries, the aim of IR systems is to process this input and to decide whether some information is useful or relevant, in order to retrieve it back to the user. Differently from Data Retrieval, which relies on well defined algorithms and data structures to allow user to exactly specify the characteristics of what he needs, IR techniques are more vague and deal mostly with natural language instead of structured data. Considering this, different strategies must be applied to approximate the user's query intention to retrievable information. The results must be ranked by their relevance, and this relevance analysis is considered to be the center of the IR field (Baeza-Yates and Ribeiro-Neto, 1999). With the growth of the World Wide Web, IR became an important discipline, since search engines became necessary to help users find relevant content on a giant network of Web pages.

Applying IR to geo referenced information (i.e. information associated to spatial descriptors such as latitude and longitude coordinates) introduces the concept of Geographic Information Retrieval (GIR). The Alexandria Digital Library project¹ is a classic example of a large collection of geo referenced materials (Smith and Frew, 1995) in which GIR was naturally required. More than just developing means to accumulate information, it was fundamental for the project to have a useful interface, including support to text-based queries. This allows users to retrieve information on what exists somewhere or where something is located, according to a variety of constraints.

¹<http://www.alexandria.ucsb.edu/>

Zhu et al. (1999) claim that the development of effective GIR techniques is one of the most challenging research issues for geospatial collections, since user queries are based on concepts and fuzzy terms, expressed by natural language. According to them, GIR systems should be able to answer two classic spatial queries: “What’s there?” and “Where’s that?”. To do this, it is necessary to describe geographic locations using both its coordinates and a set of names and textual features. Querying this textual content is a classical vocabulary problem, requiring the use of dictionaries, string matching algorithms, and toponym (place name) resolution in text (Leidner, 2007).

We must consider that some collections of information can contain items with no explicit geo references, but these items can contain information that is somehow related to some location. That is the case of the World Wide Web. Ding et al. (2000) state that a Web resource has a target audience, and this audience can be geographically enclosed. Considering this, they introduce the concept of *geographical scope*, which should point to the geographic distribution of such audience. According to them, this geographic scope is subjective, as much as relevance is to IR, since it depends on the user’s expectation and how well his intention can be expressed when querying the search engine. They propose a geographically-aware search engine, using estimation algorithms based on Web page content and HTML link information to automatically compute the geographic scope of Web pages.

2.2 Geographic Context in Documents

Buyukkokten et al. (1999) presented one of the first works on automatic identification of location information in Web pages to improve search engine ranking strategies. They mention some basic strategies that can be used. The first one is to identify geographic entities references on the page’s contents, such as place names, area codes or zip codes. Considering a search engine that knows the user’s location, the classic textual proximity approach for keyword search should still apply, using those geographic references to influence the relevance calculus, thus giving better rank positions to pages that contains places that are close to the user’s location. In practice, a user in the city of *Rio de Janeiro* querying for ‘Italian Restaurant’, would get as best results the pages that mention these keywords but also mention the city name. The second strategy is based on the analysis of the Web pages’ Uniform Resource Locator (URL) and Internet Protocol (IP) addresses. The URL could inform the country where the website organization resides (e.g. *br* points to Brazil in *www.ufmg.br*). The IP address may provide a way to find out where the website is hosted. But they point out that IP addresses

are often allocated in an *ad-hoc* fashion, and that can cause an erroneous connection between an IP address and its physical location. Furthermore, even though a website can be served in some country, it does not mean its content is limited to the the same locality.

Gravano et al. (2003) say that Web pages can be characterized as global or local, considering the interest of users in their content. But, unfortunately, at that time, search engines did not perform any kind of analysis on the *geographic locality* of queries and users, producing sub-optimal results. They propose a method to categorize queries according to their *geographic locality*, by considering two possible classes: global or local. They mention that such locality can often be implicit. Their proposal is based on machine learning tools, which use classifiers over a sample of queries from a real Web search engine, indicating the importance of further research about the geographic aspect of keyword search.

We already mentioned in Chapter 1 that Wang et al. (2005) categorize previous initiatives on GIR applied to Web pages into three main approaches. First, it is possible to obtain an approximate location of the Web server, as informed by services that relate an IP address to a pair of coordinates. Second, the location is inferred by looking at concentrations of users that access the document and their IP-determined locations, or by the location of documents that refer to it. Third, the location can be inferred by analyzing the textual content of the document. As previously discussed, this dissertation works with the last type, since the location of the Web server can be completely unrelated to the subject of the document and the IP locating techniques are, sometimes, error-prone and imprecise.

Borges et al. (2007) show that there can be many indications of geographic location in Web documents, but not all pages include unambiguous and easily recognizable evidence such as postal codes or telephone area codes. In that work, an ontology-based approach is presented for recognizing geographic evidence, including postal addresses and their components. Some other works have also focused on identifying the geographic context of Web pages by obtaining references to place names or data such as postal addresses, postal codes, or telephone numbers (Ahlers and Boll, 2008; Blessing et al., 2007; Witten and Frank, 2000), then performing some sort of geocoding (Davis Jr and Fonseca, 2007).

Silva et al. (2006) propose the identification of the geographic scope of a document using machine learning techniques. The authors, however, warn that doing so directly is hard, due to the extensive number of classes (i.e., locations) and the relatively small number of features that can be used in the classification process. Therefore, they propose a technique that first recognizes geographic evidence in text, and then use

a graph-based approach akin to PageRank (Brin and Page, 1998) to associate scopes to documents, based on a geographic knowledge repository. Such a knowledge base is essential for the process, since it contains information such as place names, postal codes, and even historical names, as provided by TGN, the Getty Thesaurus of Geographic Names² (a gazetteer).

Beyond the recognition of place names, we observe that many other terms that can occur in a text can be related to places as well. For instance, terms associated to historical events, monuments, commercial activities, names of authorities, sports teams and others can provide clear indications of geographic location, as long as the semantic connection between the term and the place can be somehow established. Such terms might be used either to establish a location or to disambiguate places that share the same name. The feasibility of this idea was explored by Backstrom et al. (2008), who present a model to track spatial variation in search queries, showing the geographic concentration of the origin of queries related to sports teams. Cardoso et al. (2008) call these terms *implicit geographic evidence*.

Considering the works mentioned in this section, we notice that methods for the geographic scope computation can be improved by the proposal of semantic-based approaches. The introduction of knowledge from other sources in the place names recognition can point to new challenges, and that is when Wikipedia starts to play an important role, as described in next section.

2.3 Wikipedia as a Source of Information and Knowledge

Recent work has shown that Wikipedia can be a valuable source of information, considering the semi-structured annotations that exist in its entries and the usual richness of articles links, which compose a *de facto* semantic network. Kasneci et al. (2009) present their approach to developing and maintaining a knowledge base called YAGO (Yet Another Great Ontology), for which knowledge sources are Wikipedia's infoboxes (sections containing attribute-value pairs) and categorical lists, enriched and automatically interpreted with the aid of WordNet³.

Cardoso et al. (2008) also use Wikipedia to experiment with named entity recognition, and present a system that can find implicit geographic references, such as determining that a text refers to New York City from the expression "Empire State Building".

²<http://www.getty.edu/research/tools/vocabularies/tgn/>

³<http://wordnet.princeton.edu>

Buscaldi (2009) use the Wikipedia as a rich source of geographical evidence and propose to build a geographical ontology considering geo-political entities found in the encyclopedia's entries. In another work, Buscaldi and Rosso (2007) compare different methods to automatic identify geographical articles in Wikipedia.

Machado et al. (2010) present an ontological gazetteer (geographic dictionary) called *OntoGazetteer*, which goes further the traditional cataloging of place names and includes geographic elements such as spatial relationships, concepts and terms related to places. Wikipedia is used to provide lists of keywords and expressions associated to places, which allows to take advantage of the gazetteer structure to improve geographic information retrieval tasks such as the disambiguation of place names in texts.

Medelyan et al. (2008) describe a method to identify topics in text using Wikipedia articles as a controlled vocabulary and as a source of statistics concerning anchor texts that are used to refer to those topics. In this same line, Milne and Witten (2008) present a text enrichment strategy that automatically adds hyperlinks to Wikipedia entries, in a process known as *automatic wikification* (Mihalcea and Csomai, 2007). This is done without any parsing or natural language analysis, but only by matching terms from the text to Wikipedia entries. Such methods allow to connect any piece of plain text to the network of concepts represented by the Wikipedia entries. It is interesting to notice that Wikipedia pages are connected to each other by hyperlinks, which demonstrates whether a concept has some importance in the context of another. Wikipedia pages also have expected sections, domain specific infoboxes, shared categories and other useful features that can inspire the proposal of automatic applications that take advantage of the encyclopedia's information. These possibilities open the way to further research on doing something else with texts submitted to the Wikification task. In this dissertation we direct this potential to the problem of the automatic computation of geographic scope.

Chapter 3

Using Evidence from Wikipedia to Automatically Relate Texts to Places

We have already mentioned in Chapter 1 and 2 that automatic applications can look for references in texts to try to capture their geographic scope. While place names are the most common element used to do so, we consider the place name to be just one of the many clues to identify the relationship between a text and a set of places. Places have characteristics such as their history, culture, social activity, economy, sports, food, folklore, arts and many others. All of those can help to decide which places are referenced by the text. Also, people are naturally bound to places, such as the city they are from, where they live and where they work. For instance, if the name of a celebrity or a public figure is mentioned in the news, we can enumerate a list of possible places related to that news. Basically, since everything exists or happens sometime and somewhere, it is quite obvious that a fact or entity will be always linked to one or more places. We should consider all of these other references as indirect evidence of the relationship between entities and places.

Just as place names need to be listed by a dictionary (a gazetteer) in order to be recognized as such in a text, indirect evidence also need some technological support to be identified. We suggest two approaches. The first one is to start with a list of target places and then list indirect evidence related to each one of them in the form of textual expressions. These expressions can be matched in text the same way we match place names. To do this, we need a good source of place-related expressions. The second approach is to use a Knowledge Base (KB) to help understand the relationship between concepts and places. Once we identify some concepts in the text, we need

ways to explore a concept network and decide which places are the most related to the text contents.

In both cases, we see Wikipedia as a suitable alternative to our proposal, since it contains a lot of textual content to be used as a source of expressions, and it also contains a network of pages connected by links, making it possible to explore the relationships between concepts. In the next sections we explain how Wikipedia can be used to automatically gather information, and the benefits of choosing the digital encyclopedia as an alternative Knowledge Base (secio 3.1). Then we present both approaches to use Wikipedia to aid the automatic identification of geographic scope in texts, one based in place-related expressions (section 3.2) and the other based on the exploration of the concept network (section 3.3).

3.1 Wikipedia as Source of Geographic Evidence

Wikipedia¹ stands out as a successful example of how a billion-user network can build something hugely useful in a collaborative way, what is called crowdsourcing. By allowing its visitors to create, edit and revise any article, something called a *wiki system*, the digital open encyclopedia has become a very popular reference in the search for concepts on several subjects. While “gathering of all world knowledge” sounds like a pretentious goal, the website’s success relies on the fact that the available content is being constantly scrutinized by million of altruistic users. Their efforts are clearly visible by the huge amount of revision that the existing articles undergo and how fast such revisions occur. According to the Wikimedia Statistics² on May 31, 2010, the English version of Wikipedia is the largest one with more than 3 million articles, more than 7 million views per hour, and an average of 23 revisions per article a month. Although the number of new articles seems to be getting stable, the website’s popularity shows to be increasing along the time, which helps to keep the collaborative philosophy in the validation and enrichment of the encyclopedia’s content.

With this constant refinement of the Wikipedia articles, there are some interesting initiatives from the community. Many of the projects proposed by users aim to normalize articles within a common domain (e.g. articles about countries, list of presidents, music groups by genre), trying to guarantee a common structure and minimal quality level in an article. To do this, they try to better wikify the text (including proper links from and to other articles) and to maintain a regular topic structure for

¹Wikipedia - <http://www.wikipedia.org/>

²Wikistats: Wikimedia Statistics - <http://stats.wikimedia.org>

similar concepts. They also include articles on their expected categories and insert useful code templates.

The categories are used to organize the articles in groups according to the branches of knowledge they belong to, and they can also have sub-categories. Such a feature allows to navigate over a taxonomy, looking for correlated concepts that share some property (e. g., people who were born in a certain year). Considering the Belo Horizonte city article in the Portuguese Wikipedia³, it is categorized inside a “*Belo Horizonte*” branch, which includes other articles about the city. On the other hand, this branch is listed under “*Capitais do Brasil*” (Capitals of Brazil), “*Municípios de Minas Gerais*” (Municipalities of Minas Gerais state) and “*Cidades Planejadas*” (Planned Cities). By exploring the Municipalities of Minas Gerais state category, we can find other sub-categories about important cities on this state. This category is also further categorized inside a more general category called “*Municípios do Brasil*” (Municipalities of Brazil), which includes links to categories of other groups of municipalities for each one of the 27 Brazilian states. If we continue doing this, exploring the categories of categories, we naturally get to more general levels, which allows to find articles on the whole world geography. Figure 3.1 illustrates this possibility with an example of Wikipedia’s geographic categories hierarchy, starting from an article about a neighborhood in Belo Horizonte city, and reaching a very general category about municipalities of the world.

We consider the categorization structure an important aspect of Wikipedia if one decides to use the digital encyclopedia as a source of geographic evidence. The exploration categories can suggest automatic or at least semi-supervised strategies to create a list of target places considered in the geographic scope computation. Without those geographic categories, the search for place-related articles would have to be done manually, by determining which Wikipedia page corresponds to each of the known places, probably with the help of Wikipedia search tool.

Code templates help to enrich the text with data in a specific format that can be parsed by the Wikimedia system. One of the most useful templates applied in common domain articles is a table of attribute-value pair called *infobox* (Figure 3.2). Every article from a music band, for example, may include an infobox with the band’s name, members, year of foundation, style, and so on. Since a lot of articles include infoboxes in their content, infobox is a powerful tool that helps the development of automatic mining tools over the Wikipedia content. Place related articles have their specific infoboxes. Figure 3.2 shows an adapted infobox from the English Wikipedia

³http://pt.wikipedia.org/wiki/Belo_Horizonte - accessed on May 22th, 2011

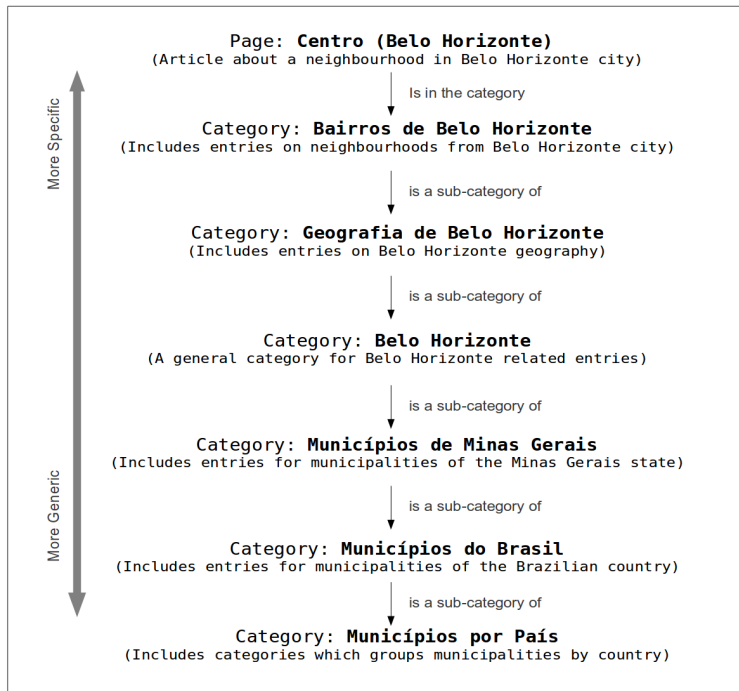


Figure 3.1. Example of Wikipedia geographic categories starting from an article about a neighborhood and reaching the categories on the World municipalities

article about Belo Horizonte. The left side of the figure shows the layout of an infobox, while the right side shows the code template used to create it. The attribute-value pairs contain information on the city name, nicknames, relation to other places, area, population and other relevant aspects.

We also consider infoboxes useful structures in the geographic analysis of Wikipedia content. They allow automatic applications to simply extract expected complex information, such as which state a city is subordinated to based on their regular structure. Since infoboxes follow a predefined code structure, they can be easily parsed. This information extraction can help applications to ensure they are analyzing the right article. For instance, when an automatic method finds the Belo Horizonte article, according to Figure 3.2, it can check if there is an infobox of the type “*settlement*”, and then check if the “*Name*” value is equals “*Belo Horizonte*”, increasing the certainty that the article is really about Belo Horizonte.

Another feature that motivates the use of Wikipedia in this work is the possibility to navigate in the encyclopedia’s network of concepts, seen as a large graph. In this graph, each node n_i represents a Wikipedia article, and an edge e_{ij} corresponds to a textual hyperlink connecting the article represented by node n_i to the article represented by node n_j . Since each article is created and maintained by the user community, we

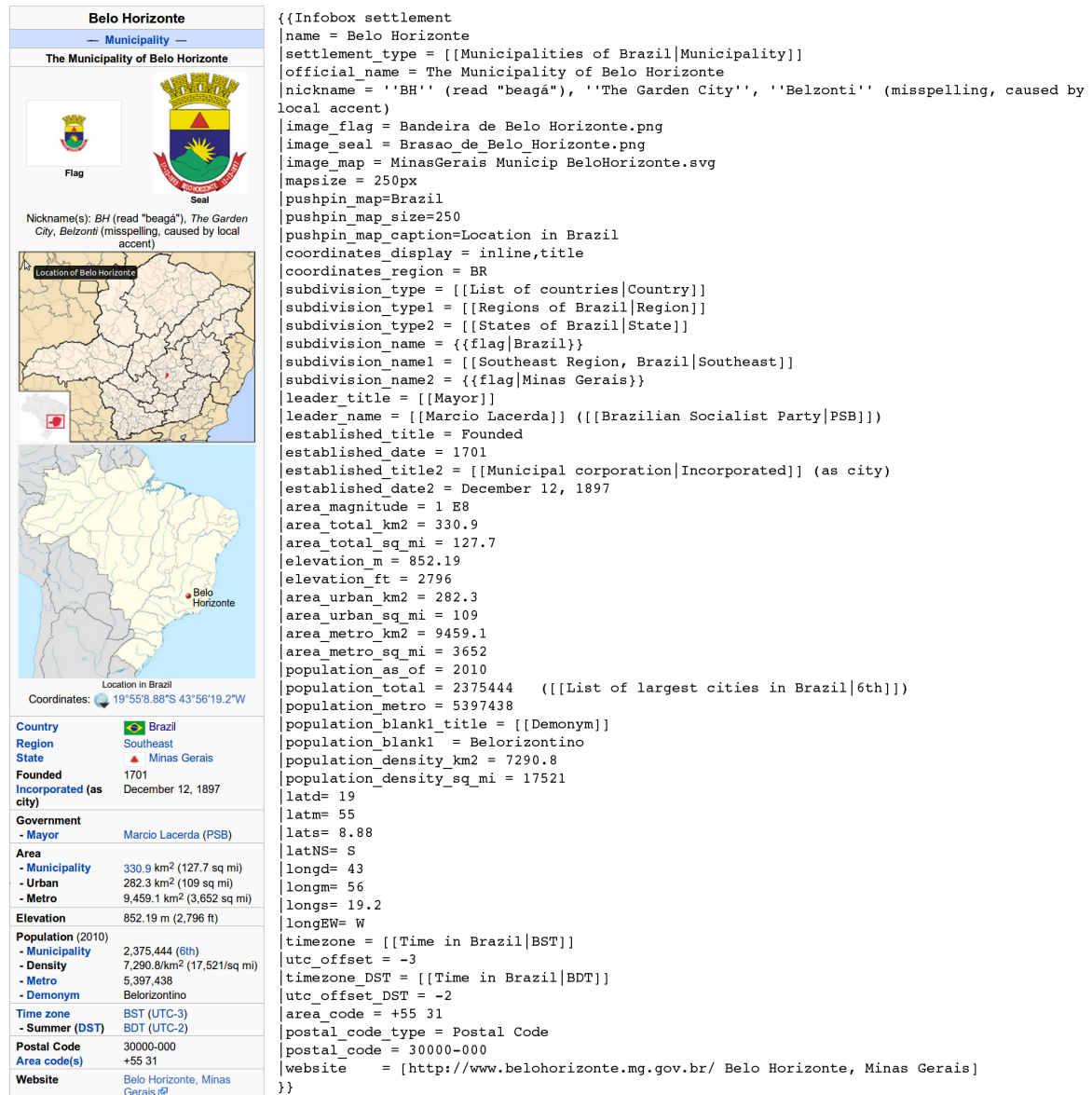


Figure 3.2. An example of infobox code template for Belo Horizonte adapted from the English Wikipedia article. On the left its layout and on the right its code template, revealing the attribute-value pairs.

assume that the links are also the subject of scrutiny, and therefore the graph structure itself is a rich network of semantically related terms and concepts. For each article, we can enumerate *inlinks* and *outlinks* sets, representing a set of concepts which refer to (in) and are referred by (out) a page. Also, the textual content has variations of anchor texts used to link to the same concept. This provide an interesting opportunity to work with alternative names or to calculate probabilistic information on whether a

concept name can vary according to the context in which it is mentioned.

If we can identify place-related articles in Wikipedia, regardless of how we do it, using Wikipedia as a source of terms and expressions associated to places becomes feasible. One can assume that Wikipedia nowadays contemplates a large set of geography-related articles containing infoboxes and organized in relevant categories. Specifically in the Portuguese Wikipedia, there are some existing community projects that focus on editing place-related articles (e. g., about cities, states, neighborhoods) creating specific infoboxes full of useful geographic aspects in each of them, and even adding alternative code templates such as Latitude and Longitude coordinates. These efforts suggest that Wikipedia can be used as a source of geographic evidence, at least for popular places. Some less popular places can have no entry in Wikipedia, which may be an obstacle. But we consider that if these places are so irrelevant that they were not listed in Wikipedia, they will probably also be irrelevant for the geographic scope computation on the Web.

Regarding Wikipedia's information quality, such content includes a lot of noise or even missing information for every article. It is not possible to ensure the information accuracy from the educational perspective. But since Wikipedia has millions of articles, it opens possibilities to develop techniques which rely on its statistical importance more than on its conceptual correctness.

3.2 Strategy 1: Identifying Geographic Context through Classifiers with Wikipedia

For this first strategy, we want to demonstrate that Wikipedia can offer textual clues to analyze texts and help in the determination of their geographic scope. To accomplish this goal, we start with a simpler problem in which the expected output is a single place name, for the place that best represents the document's scope. Then, we propose to use an automatic classification approach. Basically, we define a set of places P to be our classes, and propose a process to decide which class is more closely related to a sample text.

In order to use Wikipedia in this classification approach, we extract lists of important terms for each one of the places considered in the classification task from the encyclopedia. Those terms can be used later as a controlled vocabulary that provides discriminative information about each place in a text. We propose to obtain these terms from the hyperlinks that connect a place-related Wikipedia page to other pages inside Wikipedia. For this, we consider the Wikipedia graph perspective explained on

section 3.1, in which there are n_i representing pages, and an edge e_{ij} obtained from textual hyperlinks connecting node n_i to node n_j . As explained earlier, the connections around a node in this directed graph can be separated in two sets: inlinks and outlinks. The title of the page represented by the counterpart node in those links represents the set of relevant terms for each place.

In order to put the graph together, we downloaded a MySQL dump of the pages and pagelinks tables of the Portuguese version of the encyclopedia, obtained in Wikimedia’s downloads page on November 11, 2009. The `pages` table contains the pages themselves, including attributes such as an identifier, title, and namespace (type). The `pagelinks` table stores the connections between pages. We used the titles of entries that are adjacent to a page about some place of interest as a set of names relevant to identifying text about that place. For instance, the “Rio de Janeiro” (Brazilian state) entry includes links to pages titled “Samba” (the dance), “Carnaval” (a popular festivity), and “2016 Olympics” (an upcoming event). These titles are then included in the outlinks set. Furthermore, pages titled “French invasion” (an historical event), “Sugar Loaf” (a landmark), and “Southeast region” (a Brazilian region that includes Rio de Janeiro state), include links to the “Rio de Janeiro” entry, and are added to the inlinks set.

We consider the inlinks and outlinks to have different meanings, and potentially different values for the classification. Outlinks are found in the place entry text, thus covering names that are important for someone interested in the place. On the other hand, inlinks are links found in other entries that refer to the place, thus covering names of entries for which the place is important.

Comparing the inlinks and outlinks term sets related with each place in P , it is expected that some terms occur in relation with more than one place. The terms that are related to a smaller number of places in P are considered to be more discriminative than others that relate to many places. For instance, the term “Southeast region” relates not only to “Rio de Janeiro”, but also to “Minas Gerais”, “Espírito Santo” and “São Paulo”, the other states in the same region, and can occur both in the inlinks and in the outlinks sets of each of these states.

To work with automatic text classification, we have to represent each text as an array of features represented by numeric values. A basic approach would be to define those features as binary values indicating the presence or absence of some words or terms from a vocabulary. Advancing this model, we can also define those features as words frequency, giving more information for the classification algorithm.

We propose a simple measurement of the discriminative power of terms, to be used as weights for the classification, hereafter called $w(t)$. The number of places

from P adjacent to the term t in the graph is divided by m , the total number of places considered as classes. This ratio is then normalized and squared, so that less discriminative terms get a weight that is close to zero, while more specific terms get weights closer to the maximum of one. For instance, if we use the set of 27 Brazilian states as our target classes, a term t_1 that is related to a single state obtains a weight $w(t_1) = 1.0$, while a term t_2 that is related to 26 of the 27 states gets $w(t_2) = 0.0054$. Equation 3.1 shows the formula for $w(t)$, in which $adj(t)$ represents the number of places from P that are adjacent to the term t in the graph. Naturally, the value of $adj(t)$ is at least 1, since t only makes part of the list of terms if it is found in at least one link from or to a place in the Wikipedia graph.

$$w(t) = \left(1 - \frac{adj(t) - 1}{m}\right)^2 \quad (3.1)$$

A basic classification strategy would determine the frequency of the terms in the text, and perform a weighted sum using $w(t)$. The place that reaches the highest total would be the result of the classification. However, in a preliminary manual investigation, we verified that this simple strategy could produce many distortions. First, no distinction between terms from inlinks and from outlinks could be made. Second, some places could be penalized in the classification, because of their lower popularity or because of a less detailed Wikipedia entry. New normalization schemes would then be necessary, potentially leading to difficulties in the analysis and generalization of the results.

To overcome this problem, we used the Multinomial Naïve Bayes automatic classifier, known for obtaining good results in text classification using a representation based on the frequency of a given set of terms (McCallum and Nigam, 1998). The Naïve Bayes classifier operates by adjusting a model that calculates the probability of a class to generate an instance considering the presence of features in the examples. It considers each feature to be independent of the others, which is a naïve assumption, but which in practice simplifies the learning process while still generating good results. The multinomial variation of this classifier is widely used in text classification and it assumes that features represent the frequency of terms, ignoring the position of the terms in the text and only observing their occurrence (McCallum and Nigam, 1998).

The idea is to let the classifier adjust itself to the combination of quantitative information on the occurrence of terms from the Wikipedia in the text. Weighted sums $S_{i,j}$ of the occurrence of terms related to the place p_i in document d_j found among the inlinks and among the outlinks are fed to the classifier as two separate sections of the sets of features. Equations 3.2 and 3.3 are used to determine the value of each single

feature corresponding to a given place (p_i) and document(d_i) association.

$$S^{in}(p_i, d_j) = \sum_{k=1}^{|in|} (w(t_k) \times Frequency(t_k, d_j)) \quad (3.2)$$

$$S^{out}(p_i, d_j) = \sum_{k=1}^{|out|} (w(t_k) \times Frequency(t_k, d_j)) \quad (3.3)$$

By using the idea of two separate weighted sums for each place, a first classification model was then produced. In this model, a term with a high $w(t)$ adds significantly to the sum, while terms with low weights can also contribute to the result, but with a smaller impact. We generate $2m$ features per document in the collection, i.e., twice the number of places in P . Each of these features represents the relationship between a list of terms from Wikipedia (from the inlinks or from the outlinks sets) and a document. With that information, the classifier is expected to find the best fit between each document and a place from P .

Since we are also motivated to use machine learning classifiers in order to fine tune a frequency-based information model, we decided to improve the proposed classification model to generate a second version, using more detailed features. Our hypothesis was that adding more features to the model could increase the possibilities for the classifier to adjust itself to the model. Thus, we exploded the two weighted sums for each place into $2m$ weighted sums. Each set of terms (from the inlinks and outlinks) is divided into m sets, one for each possible value of $adj(t)$. With this, we obtain m weighted sums corresponding to groups of terms with equal discriminative capacity. Therefore, we put together a feature set consisting of $2m$ sums for each place in P , i.e., $2m^2$ features per document.

As an example of this second classification model, consider a simplified version of Wikipedia’s graph, including entries on two Brazilian states (Figure 3.3). The inlinks and outlinks are obtained from the text of the entries, and then weights are calculated according to Equation 3.1 (Figure 3.4).

For the classification, consider the documents shown in Figure 3.5. The terms that occur in either the inlinks or the outlinks sets for each place are underlined. Some terms, such as “Southeast”, appear in more than one list, for more than one state, and therefore are less discriminative when they appear in a document.

The first classification model then generates the features for the classifier, as shown in Table 3.1. The table shows the results of the weighted sums for inlinks and outlinks in both states. Since there are only two states, and this strategy only generates two features per place, there are four features in the table for each document. The

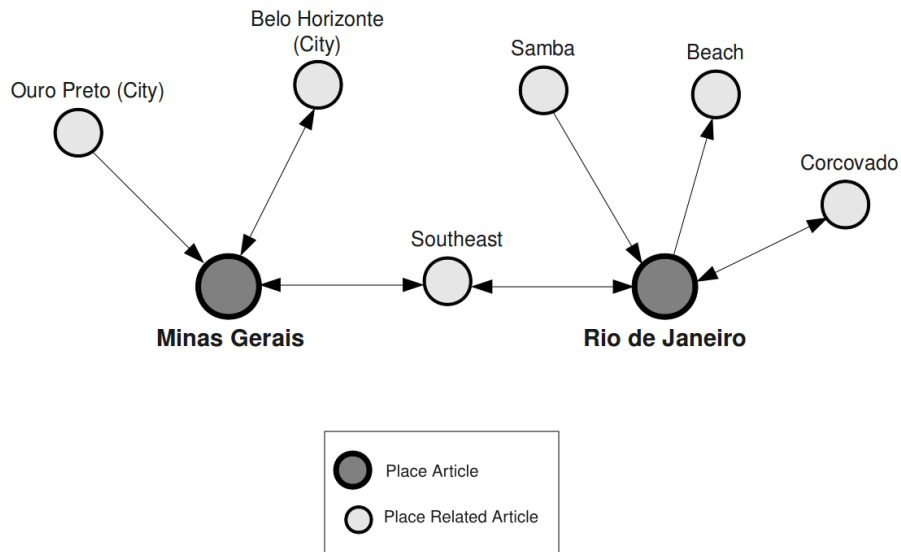


Figure 3.3. Wikipedia graph example

Minas Gerais inlinks		Rio de Janeiro inlinks	
Article	$w(t)$	Article	$w(t)$
Ouro Preto (city)	1	Samba	1
Belo Horizonte (city)	1	Corcovado	1
Southeast	0.25	Southeast	0.25

Minas Gerais outlinks		Rio de Janeiro outlinks	
Article	$w(t)$	Article	$w(t)$
Belo Horizonte (city)	1	Beach	1
Southeast	0.25	Corcovado	1
		Southeast	0.25

Figure 3.4. Term weights in the inlinks and outlinks sets for each place

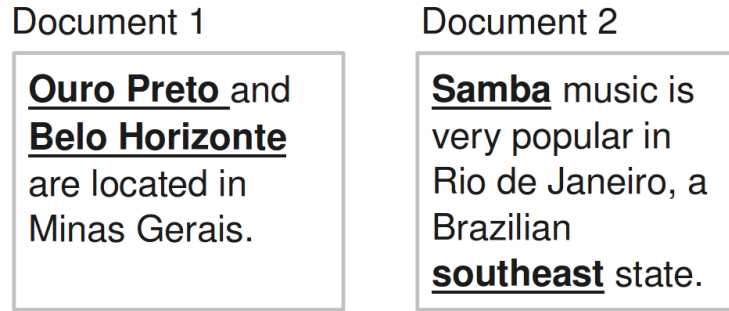


Figure 3.5. Sample documents for classification

Documents	Minas Gerais		Rio de Janeiro		Class
	<i>in</i>	<i>out</i>	<i>in</i>	<i>out</i>	
Document 1	2	1	0	0	Minas Gerais
Document 2	0.25	0.25	1.25	1.25	Rio de Janeiro

Table 3.1. Features in the first strategy, and calculated classes for each document.

classifier uses these features to decide on a class for each document. In the example, document 1 is classified as “Minas Gerais”, and document 2 as “Rio de Janeiro”. The mention to “Southeast” in document 2 generates the 0.25 values in the features associated to Minas Gerais, but the same happens in relation to Rio de Janeiro. The results for document 2 are more decisively influenced by the occurrence of “Samba”, a term that has been exclusively associated to Rio de Janeiro.

In the second classification model, terms are split in groups according to their $adj(t)$ value, which indicates how specific each term is. Since there are only two states, $adj(t)$ can be either 1 or 2: “Ouro Preto” and “Samba” have $adj(t) = 1$, while “Southeast” has $adj(t) = 2$. The weighted sum calculations are now performed separately for each group of terms with the same $adj(t)$ value. Table 3.2 shows the results for the example. Notice that the 1.25 value for the inlinks to Rio de Janeiro was split into a in_1 value of 1.0, corresponding to “Samba”, and a in_2 value of 0.25, obtained from “Southeast”. Considering such division of terms into groups, the calculation of the features still follows Equations 3.2 and 3.3, but this time the summation is performed over groups of terms that have the same discriminative power.

In order to check if this feature expansion strategy brings benefits, we performed a preliminary test of the classification models, using a set of 200 documents and two places as classes. With the first model, The Multinomial Naïve Bayes classifier was able to accurately classify 92.7% of the documents (average precision in a 10-fold cross

Documents	Minas Gerais				Rio de Janeiro				Class
	in_1	in_2	out_1	out_2	in_1	in_2	out_1	out_2	
Document 1	2	0	1	0	0	0	0	0	Minas Gerais
Document 2	0	0.25	0	0.25	1	0.25	0	0.25	Rio de Janeiro

Table 3.2. Features in the second strategy and classification results

validation (Witten and Frank, 2000) set of tests). With the second and more detailed classification model, the Multinomial Naïve Bayes classifier reached an average precision rate of 98.0% with 10-fold cross validation. The improved precision shows that having more evidence to support the classification can lead to better results, as expected. In the results section we show a larger experiment on this strategy, involving a larger collection and a larger number of classes, and a comparison to a text classification that uses no geographic evidence from Wikipedia as a baseline.

We can see that Wikipedia can be used as a source of terms for the geographic scope computation, but as the results will show in section 4.2, the representations of features as terms frequency and the machine learning approach can face difficulties, specially regarding the scalability when considering a huge number of classes. This motivates us to propose the next section strategy which, among other benefits, is not based on machine learning.

3.3 Strategy 2: Identifying Geographic Context through Topic Indexing with Wikipedia

We demonstrated on the previous section that Wikipedia can offer textual clues for the geographic scope computation. We now advance on this problem by considering that documents can be associated to multiple place names. Since machine learning process presented some issues as section 4.2 shows, and requires a training step, we now try to avoid its usage and focus on a more scalable strategy. So, this section describes a technique for tagging text documents with the names of one or more places that describe its geographical scope, considering their content. Our proposal relies on a topic indexing method that uses Wikipedia articles as expected topics (Medelyan et al., 2008). By identifying those topics in a document, it is possible to connect it to Wikipedia, used here as a semantic network. Then, we can evaluate how the document is related to previously defined target places, which can also be obtained from Wikipedia, now seen as a source of geography-related content. We also introduce the concept of *acceptance*: we can avoid tagging documents in which there is insufficient

evidence of their relationship to any place.

Our proposal consists in the steps that follow:

- **Defining a list of target places within Wikipedia:** Our method needs to know a list of places that are expected to be used as tags. Those places must have their Wikipedia articles identified. As we explained in section 3.1, Wikipedia offers some features that make it possible to perform mining for places, such as the categories system. Some works in the literature focus specifically on this task (Buscaldi and Rosso, 2007; Buscaldi, 2009).
- **Topic Indexing with Wikipedia:** In order to connect a text document with Wikipedia’s semantic network, a topic indexing algorithm must be applied, considering as a controlled vocabulary the titles of articles from Wikipedia. Such a mechanism also must offer high precision and be prepared for disambiguation, since some portions of the text can be used to refer to more than one article. We apply the methods presented by Medelyan et al. (2008) and by Milne and Witten (2008).
- **Navigating in Wikipedia’s graph to discover places:** By identifying a subset of Wikipedia articles as topics from the text, we can consider the evaluated document as a new node on the encyclopedia’s graph. The whole set of articles from Wikipedia compose the other nodes from this graph, while the internal links between each one of them give us the edges. We propose to perform a breadth-first search on this graph, starting from the new node (the text). Thus, we can meet some of the target places listed before in this search, and then we can calculate their semantic distance to our start node and their relevance in the network. This step will generate a list of candidate places.
- **Evaluating certainty and choosing responses:** Given a list of candidate places, we must evaluate if there’s enough information to choose one of them as a geographic tag for the document. A rate threshold can be used to choose the best place from the list, but in some cases the ratings of the places can be too similar, characterizing a confusing state. Many different approaches can be tested in this step in an experimental way.

The presented steps are detailed in the next sub-sections.

3.3.1 Defining a List of Target Places within Wikipedia

Considering the rich geographic content offered by recent releases of the Wikipedia, it is possible to conceive some strategies to define a list of target places to be used in the geotagging process. In the case of Brazilian places, it was possible to manually navigate through the categories system looking for those that list only geographic entities, such as Brazilian states, cities in a state, micro-regions from a state, and even more detailed categories such as city neighborhoods and boroughs. This hierarchical organization allows to write scripts to gather all entities listed from each one of these categories, creating a kind of alternative gazetteer in which entities are linked to Wikipedia articles. To avoid listing entries that do not correspond to places during this automatic collection, it is also possible to parse the content of each one of them, trying to find infoboxes that help to determine whether the entry's content is really about the desired type of place.

For our experimental evaluation in this paper, we have considered only Brazilian cities and states. After collecting all of them, we compared the number of cities per state with a official list of from the Brazilian Institute of Geography and Statistics (IBGE). We were able to cover 5514 cities, that is about 99% of the 5561 existing ones.

3.3.2 Topic Indexing with Wikipedia

Medelyan et al. (2008) described a method to identify topics in text using Wikipedia articles as a controlled vocabulary and a source of statistics concerning anchor texts that are used to refer to certain topics. This technique not only allows to identify the subjects in text but also to connect a text document with the Wikipedia articles network. Then it is possible to propose several heuristics to capture the context of textual documents. In our case, we focus on the geographic aspect and how the text relates to geographic entities described on Wikipedia.

According to the selected topic indexing method, in a first stage all word n-grams (possible sequence of words) from the text must be extracted. Then the probability of a n-gram a to be a topic (keyphraseness) is calculated by counting how many Wikipedia articles contain that n-gram as an anchor text (D_{Link}), and how many articles contain that n-gram at all (D_a). According to the formula 3.4, keyphraseness provides a value between 0.0 and 1.0 that can be used to establish a threshold to select n-grams to be included in the topics list.

$$Keyphraseness(a) = \frac{count(D_{Link})}{count(D_a)} \quad (3.4)$$

After that, if a n-gram always points to a single Wikipedia article, then it is chosen to be a topic. Otherwise, if a n-gram has been used to refer to many different articles, we must disambiguate it. This is done by calculating the commonness of an candidate article to a given anchor text, and also by calculating the semantic similarity of each candidate article to other articles already chosen as topics for the text. Commonness is calculated by equation 3.5, considering how many times an anchor text a is used to refer to an article T , and how many times it occurs at all. Semantic similarity between two articles x and y can be calculated by equation 3.6, considering their hyperlinks sets (X and Y), these set of overlapping links ($X \cap Y$), and the total number of articles in Wikipedia (N). According to Medelyan et al. (2008), based on the latter two equations, a score is calculated for each possible article, according to the equation 3.7, considering C as the set of candidate articles for the anchor text a in text T .

$$Commonness_{a,T} = \frac{P(a|T)}{P(a)} \quad (3.5)$$

$$Similarity_{x,y} = \frac{\max(\log|X|, \log|Y|) - \log|X \cap Y|}{N - \min(\log|X|, \log|Y|)} \quad (3.6)$$

$$Score(a, T) = \frac{\sum_{c \in C} Similarity_{T,c}}{|C|} \times Commonness_{a,T} \quad (3.7)$$

Also on the disambiguation process, Milne and Witten (2008) go one step further and balance the score equation by applying machine learning. The quality of the context is also considered, along with similarity and commonness. Those three features are then used to train a classifier that is shown to be able to distinguish valid articles (candidates) from irrelevant ones. A subset of Wikipedia articles is used to train the classifier that learns with the examples how to best select topic candidates on texts.

In order to apply this topic indexing technique in our proposal, we developed some Java code on top of the the Wikipedia Miner Toolkit, an open source implementation of the method presented by Medelyan et al. (2008) and Milne and Witten (2008) that is maintained by the authors team at University of Waikato (New Zealand). By using this toolkit, we were able to preprocess Wikipedia's content, creating an accessible database with all necessary information on anchor text counts used by the equations. The toolkit was also used to mine Wikipedia for place-related articles, according to our proposal.

3.3.3 Navigating in Wikipedia’s Graph to Discover Places

Wikipedia can be seen as a huge graph in which articles represent nodes and links among them represent directed edges. Since the result of the topic indexing task described in the previous section is the subset of Wikipedia articles that are topics from the text, we can consider our target text document as a new node, in which topics are the edges that connect it to the existing graph. In Wikipedia, links are included to connect entries to other concepts that help readers understand the article, correlating it to a wider context. Considering this, Wikipedia’s graph represents a large semantic network, containing several important concepts defined and refined by the common sense of many users.

Our method is based on a breadth-first search on this graph, starting from the node that represents the document being geotagged. This kind of graph search begins at a specific node and visits all of its neighbors, then visits all neighbors from the already visited nodes, and keeps going on recursively until it visits the whole graph component. During the visit to a node, we check whether it corresponds to a known place from the predefined list and, if so, we consider it to be a candidate geotag.

Considering the moment in which a place is reached in the breadth-first search, we can take the current depth of search as the minimum distance from the start node to the place. We then use this information to establish a distance metric (D_p) that quantifies the impact of a certain place p when it is found in the search. If some of the places are already present in the topics of the text document, those have $D_p = 1$. If a place is found because it is adjacent (in the graph) to some of the original topics in text, it has a $D_p = 2$.

Concerning the search depth, we use a limit (*maxdepth*) in order to avoid the full search of the graph. This must be done because if we explore the entire graph component that is accessible from our start node, we will probably get much unnecessary and confusing information, distant from the original context. Notice that when we find a few places in the first depth levels, they probably contain edges to more general places (e.g. states have links to their corresponding countries). If we keep searching, those general places will also link to other specific nodes, siblings of the ones found first, and that could decrease the chance of finding relevant candidate places. In preliminary tests we observed that a *maxdepth* of more than 2 levels had covered a large portion of nodes from the almost 1 million nodes of the Portuguese Wikipedia graph. In the result section 4.3 we present more details on this issue.

Figure 3.6 shows the related places discovered on the Wikipedia graph. In the first moment (1), there is a sample text as a root node in this search, and the found topics

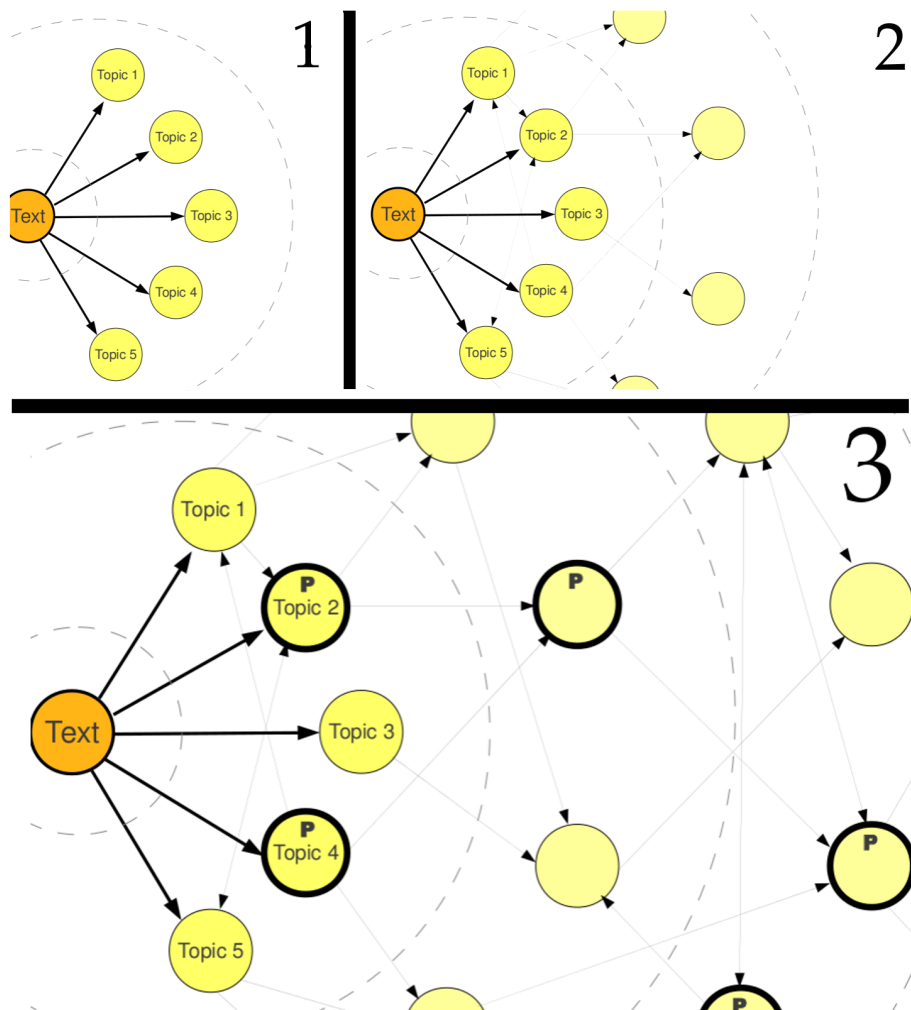


Figure 3.6. Discovering places related to a wikified text in Wikipedia's graph.

as the first connections to the Wikipedia graph. In the second moment (2), we start to work with links between those topics and from the topics to new Wikipedia entries, representing other nodes from the graph. In the third moment (3), after exploring the Wikipedia graph in deeper levels, we can look for those nodes that are about places to proceed with the geotagging process.

We also propose in our method a metric called *adjacency count* (C_p) of a place node p . During the search, every time we visit a node we need to get a list of all its adjacent nodes in order to feed a queue of nodes to be visited later. If a place node is in this adjacency list, we increase its adjacency count. This metric informs us the relevance of some place to the sub graph obtained at the end of the search. If many topics from the text document are strongly related to a certain place in Wikipedia's graph, then its corresponding node p will probably get a high C_p value, since many

nodes point to it.

The distance and adjacency count metrics for a place p can be combined to generate a final metric called $Score_p$, which decreases the C_p value dividing it by D_p to the power of an *exponential decay* factor (See equation 3.8).

$$Score_p = \frac{C_p}{(D_p)^{expdecay}} \quad (3.8)$$

As a result, C_p represents the relevance of the place in the context of a document, but this the $Score_p$ degrades as this place is considered to be farther away from the original topics in the semantic network. As a final result from the search step, we obtain a list of candidate places P , each of which accompanied by a measure of the relevance of the place to the context of the textual document we want to geotag.

3.3.4 Evaluating Certainty and Choosing Responses

Given the list P of candidate places, we must now generate a set of tags that define the geographic scope of the text document. First of all, if P is empty, it means that the topics identified in the text could not provide any information on the geographic entities related to the context. In this case the document is not geotagged.

If one or more places are listed in P , $Score_p$ provides a filtering measure, i.e., $Score_p$ allows to select the most relevant tags, and filter out less relevant responses. We propose some selection strategies, with varying selectiveness:

- **Top k :** sort tags downward by $Score_p$, then select the first k tags.
- **Global Threshold:** define a minimum value v for $Score_p$, and return all p in P where $Score_p \geq v$. $Score_p$ must be normalized in order to define a general-use, percentage-based threshold.
- **Relative to First Threshold:** considering the candidate places list P sorted downward by $Score_p$, calculate the percentage of gain g relative to the first place p_1 to the next places p_i using $g_i = Score_{p_1}/Score_{p_i}$. Define a minimum gain value v and return all p_i so that $g_i \geq v$. The idea is that the selected tags must be nearly as good as the first one in the ranking. Low g values allow less important tags to be included in the response, while values close to 1.0 make it return only those tags in which $Score_p$ is close to the top one. At least the top scorer will be part of the response.
- **Relative to First Threshold with Top- k :** the same as the previous strategy, but limiting the number of responses to the first k ranked.

The next chapter presents experiments to evaluate both strategies presented in this chapter. Regarding the topic indexing strategy presented in latter sections, a collection of news associated to Brazilian states is geotagged using the proposed method. Then we discuss the results in order to confirm the feasibility of our proposal and determine the scope for improvement.

Chapter 4

Experimental Evaluation

Now we present an experimental evaluation of both strategies proposed in the Chapter 3. The tests are all based on a collection of news we have built, in which every single news is associated to a Brazilian state as its geographic scope. We characterize this collection in section 4.1. Then, in section 4.2, the automatic classification strategy proposed in section 3.2 is evaluated. We demonstrate the feasibility of using Wikipedia to generate features for automatic text classification using machine learning. We use as baseline a classic automatic text classification task that does not use our Wikipedia features scheme, and then we check our strategy lacks and advantages. Finally, in section 4.3, we experiment with the topic indexing with Wikipedia strategy presented in section 3.3. Automatic evaluations are performed based on the precision of the method when trying to geotag news with the right state. We also perform a manual evaluation, with a small subset of the news collection, in order to better understand the circumstances that lead the process to some of the errors revealed by the experiments.

4.1 Building the Test Collection

For each of the 27 Brazilian states, we performed a manual search for relevant news websites about the locations. In those websites we usually find a section dedicated to local subjects, and we considered it to be a good source of texts with a geography scope limited to a single state level. We attempted to collect about 100 news articles for each state, and we read their titles before collecting them in order to be sure they were really about a local subject. Only the title and body text of the articles were stored, using an automatic collector based on XML Path Language (XPath) expressions. As a result we built a collection of almost 2700 text documents in Portuguese, all of them with a single Brazilian state as an expected label (see Table 4.1).

The news sources was chosen based on the presence of a local section in the news website, and also by observing how easily the news could be collected by XPath expressions. The relevance and popularity of such sources was not the main aspect observed, since the goal was to collect texts strongly related a certain place, not matter where they were found.

For the automatic classifier strategy (Section 3.2), we used only a subset of the collection, containing only the following states: Minas Gerais, Rio de Janeiro, São Paulo, Acre, Amazonas, Bahia, Santa Catarina and Pernambuco.

For the topic indexing strategy (Section 3.3), we used the whole collection. The topic detection algorithm presented by Milne and Witten (2008) was adapted to the Portuguese version of Wikipedia using a XML dump of the digital encyclopedia released in March 2010. For caching purposes we performed the topic indexing process over all documents from the collection. The topics were stored in text files as a list of Wikipedia pages IDs. Both data sets, the raw news text and the cached news topics, are available at our laboratory's website ¹ for public usage.

4.2 Results for the Classifier Strategy

In order to check the validity of the automatic classification strategy (Section 3.2), we performed experiments on the single label classification of news documents according to a subset of the Brazilian states. Test data were extracted from local or state news sections of newspaper Web sites. We then performed an automatic classification using a bag-of-words approach with a TF-IDF weighting scheme to serve as a baseline. Finally, we applied the proposed method to the same documents and compared the classification results to the baseline.

After collecting and extracting news texts, each document had their stopwords removed and their remaining terms stemmed, so that the classification process would only work with keywords. We used the stopword list for Portuguese compiled by the Snowball project². Regarding stemming (a process that reduces words to their radical form, eliminating variations such as plurals, gender, and verb tenses), we used the Orengo algorithm for Portuguese (Orengo and Huyck, 2001), as implemented in the PTStemmer project³. The same treatment was applied to the titles of entries from Wikipedia, in order to allow the matching of terms in the documents. We only considered news from 8 of the 27 Brazilian states, including three important ones

¹<http://www.lbd.dcc.ufmg.br/collections>

²Snowball, <http://snowball.tartarus.org/> - accessed on Dec 3 2009

³PTStemmer, <http://code.google.com/p/ptstemmer/> - accessed on Dec 3 2009

(Minas Gerais, Rio de Janeiro and São Paulo), and others which are representative of the other geopolitical regions of the country.

4.2.1 Baseline: Traditional Automatic Text Classification

The usual technique for text classification is based on comparing term distributions and occurrences in a document to the frequency of terms that are common to previously defined document classes. The semantics of each term is irrelevant for this technique, since the classifiers only compare terms as strings of characters. We used this approach to classify documents from the test collection in order to get a baseline result.

The popular bag-of-words model reduces documents to lists of terms, presented to the classifier as a set of TF-IDF (Term Frequency, Inverse Document Frequency) measurements (Salton and McGill, 1986). In this model, each document is considered to be a set of terms, and a union set of terms is built from all documents. The TF-IDF measurement is calculated from two components. The TF component represents the frequency of the term in a document, normalized by the number of terms present in the document (Equation 4.1). The IDF component measures the inverse frequency of the term in the collection, log-normalized by the number of terms in the union set (Equation 4.2). The final measurement is the product of the two components for each document (Equation 4.3). TF-IDF results are low for common terms, and high for rare terms.

$$tf_{i,j} = \frac{tf_{i,j}}{\sum_k n_{k,j}} \quad (4.1)$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in D\}|} \quad (4.2)$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (4.3)$$

Several classifiers could be used with this document representation. We chose Multinomial Naïve Bayes Classifier due its popularity, simplicity, and ease of operation to obtain good results when working with term frequencies (McCallum and Nigam, 1998). Using TF-IDF, more than 4,000 features were required to represent the same small collection of 200 documents used in preliminary tests. In that same test, the accuracy for the Multinomial Naïve Bayes classifier using TF-IDF was 97.08%, a result that reaffirms the high efficiency of this type of classification, as verified by previous works (McCallum and Nigam, 1998). This result is used as a baseline for tests using our method, as shown in the next section.

4.2.2 Experiments and Results

Experiments were performed using Weka (Witten and Frank, 2000) version 3.6.1, a data mining tool that includes the Multinomial Naïve Bayes classifier. Java programs were developed to extract features from the document collection, producing datasets in Comma Separated Value (CSV) format for each of the models: Wikipedia evidence and TF-IDF. In these experiments, using the single-label approach means that each text will be related to a single state, even though news sometimes refer to more than one state. Our goal is mainly to show the validity of the Wikipedia-based approach, and for that the single-label classification is adequate. Associating a text with multiple states can be achieved by establishing thresholds for the weighted sums, so that the text can be considered as associated to all places for which $w(t)$ exceeds a given value. Testing this alternative is reserved for future work.

The tests that follow compare the final Wikipedia evidence classification model presented in Section 3.2 to the TF-IDF model presented in section 4.2.1. First, we performed a test on the capacity for each model to succeed with varying training set sizes. We divided the total of 831 documents in two balanced parts, each one of them holding half of the documents of each class. The first part was reserved to be our test set. The second part was used as a source for training sets. From it, training sets in various sizes (100%, 80%, 60%, 40%, and 20% of its original size) were generated by selecting random instances and generating a sample with the proportional distribution of each class.

We ran our two classifiers over those training set sizes testing the resulting model against the same dataset. This preliminar test allowed to verify how much training data is necessary for each classification strategy to achieve successful results. Figure 4.1 and Table 4.2 show the results of this test.

Notice that the classification strategy using Wikipedia evidence starts with a accuracy close to 60%, and reaches more than 75% with only 60% of the training data. As shown in Section 3, instead of working with the bag-of-words feature model, our method groups term frequencies in general features considering the place, the adjacency type, and the discriminative level obtained from Wikipedia links. As a result, a relatively small amount of features is used to represent the document, thus allowing the Multinomial Naïve Bayes classifier to generalize the model with fewer training instances.

On the other hand, the TF-IDF classification strategy starts at about 12% of success, increasing as the size of the training set grows. The accuracy only approaches the one obtained using Wikipedia evidence when all of the data are used for training.

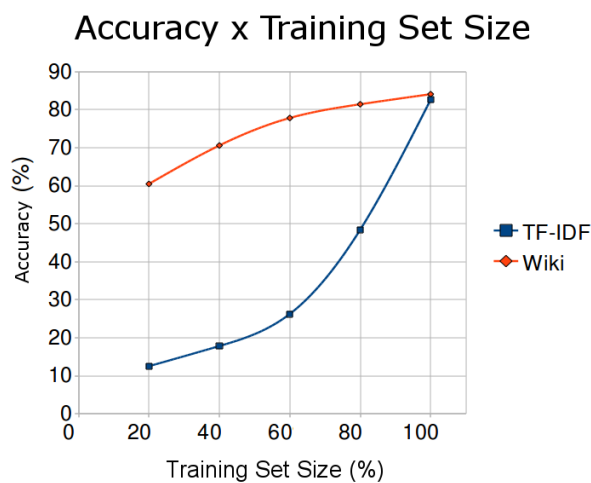


Figure 4.1. Accuracy versus training set sizes

The gradual increase from a low rate of success shows that the classifier was unable to generalize much for unknown cases, reaching adequate results only with the totality of the training set. This may be explained by the fact that, when using a bag-of-words approach, many features are required to represent the documents. Thus, when training is performed with only a few documents, many rare terms that contribute to the characterization of each class are probably not in the sample, reducing the necessary information for generalization. Therefore, the Multinomial Naïve Bayes classifier is unable to adequately adjust the probability for some terms to indicate the correct class for some documents.

A second evaluation was performed, this time varying the number of classes used for classification. Seven different datasets were generated, ranging from two to eight Brazilian states. States were added to the datasets in the following order: Minas Gerais, São Paulo, Rio de Janeiro, Acre, Pernambuco, Bahia, Santa Catarina and finally Amazonas. For both classification models, 5 executions with different random seeds (meaning different criteria for the random selection of the training and of the test sets) were executed, using the 10-fold cross validation technique in each execution (the dataset is divided into 10 parts, and in each run 9 parts are used for training and the remaining part is used for classification testing). This technique ensures that every document is classified, and the final accuracy is the average result of runs with 10 different training sets. Notice that, in the previous test, the methods achieved similar results with large training sets, thus allowing for a fair assessment as to a varying number of classes. Figure 4.2, Table 4.3, and Table 4.4 show the results.

Notice that the classifier based on Wikipedia evidence showed good results for

few classes, but its performance degrades as new classes are added. The decreasing rate is of about 5% for each new class. However, increasing the number of classes implies in adding 54 new features for each new class, and therefore the complexity of the model grows as new classes are considered. The TF-IDF classifier showed some instability in the results, suggesting a behavior close to 88% with 4 classes or more. Using all 8 classes, the TF-IDF classifier overcomes the proposed alternative by 4%, although the bias of this result is going to be discussed on the next tests. Both models showed low standard deviation. The TF-IDF approach has a lower region in the graphic (figure 4.2) that probably is the result of a classifier confusion with the documents included in the second increment (from Rio de Janeiro), because they have the same source as the second class documents (from São Paulo) (See Table 4.1). To check such confusion, we have included the documents from São Paulo and Rio de Janeiro classes in different orders, and we could see that the accuracy decreased every time both classes are present in a dataset. The gradual decrease of the Wiki curve is probably associated to the increasing probability that a term is adjacent to more than one place as more places are included.

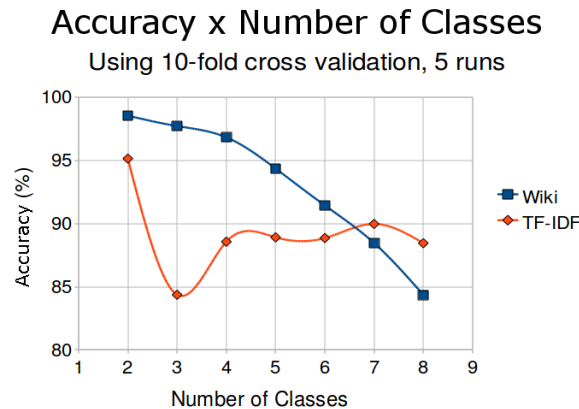


Figure 4.2. Average accuracy versus number of classes

The next tests analyze an aspect in which the Wikipedia evidence approach shows good results in comparison with the bag-of-words TF-IDF method. Even though the model based on the Wikipedia produced results below the baseline in some circumstances, it is important to observe that the use of TF-IDF does not ensure that the classification actually takes into account geographic factors on the document contents. Some text elements that are common to each news site, or expressions that are preferred by reporters of the same source may somehow influence the TF-IDF results. In other words, the classification may be based towards classifying the writing style of

each source instead of their geographic scope. Discriminative terms that do not represent geographic evidence are ignored in the Wikipedia-based strategy. Therefore, the success of the TF-IDF classifier in some situations does not ensure that the identification of geographic scope of documents would be accomplished successfully for other types of documents.

In order to verify such geographic bias of each classification method, a third test was performed. We assembled a list of place names related to each of the 8 Brazilian states that are included in this test collection. This list includes the name of the state, its acronym, the names of the 10 most important cities, some touristic sites, popular names for regions, and others. Over 100 terms were listed. Then, the collection was pre-processed, removing names in the list from the documents. More than 35,000 removals were made. The modified collection was then reprocessed by both methods, again using the Multinomial Naïve Bayes using 10-fold cross validation. We expected the impact of the removal of geographically-significant terms to be greater for the method that actually relies on place-related evidence.

Results showed that the TF-IDF method obtained a accuracy of 82.4%, i.e., about 6% less than the previous result. The Wikipedia-based method obtained a accuracy of only 56.0%, thus suffering an impact of more than 30% from the removal of place-related terms. We conclude that the TF-IDF method, in spite of the good classification results, does not ensure that geography is actually being considered in the classification. Therefore, we hypothesize that tests using a wider variety of sources would cause the TF-IDF method's accuracy to drop. Verifying this hypothesis is left for future work.

To extend the conclusion of the third set of tests, another experiment was performed. In order to verify the low performance that is expected of the TF-IDF classifier for the identification of the geographical context of documents that comes from a different source, a small test collection was created using the text of the Wikipedia articles for each one of the 8 Brazilian states. The training set was composed of the same initial 831 news documents used in previous tests. In other words, the computer learns with the set of news and tries to classify unknown texts very distant from these news. The Wikipedia articles are rich in contextual geographic evidence, so good results are expected if the classifier is sensitive to geography-related terms. Results showed that, as expected, the Wikipedia-based method achieved a 100% accuracy, while TF-IDF achieved only 50%. Of course, the test is biased in the case of the Wikipedia-based method, since the classification evidence includes the entry titles obtained from Wikipedia's graph, which must be a part of the text of the test set. However, this test shows that traditional text classification is clearly outperformed by the method proposed in this dissertation. While our method gets better results in

the presence of geographical evidence, the traditional method tends to improve in the presence of any kind of textual evidence.

These results confirm that Wikipedia keywords and expressions can be used for geographic scope computation, but some of the problems we have noticed motivated us to try different methods. The use of machine learning techniques clearly presents some limitations when we start to consider a more complex situation, and that confirms the conclusions of Silva et al. (2006). Both, the Wikipedia and the TF-IDF model present some issues on how the number of features grows. Wikipedia approach needs more features when we add more classes, since it needs a determined number of features for each place. TF-IDF model increases the number of features when we classify a larger collection of texts. More texts means more keywords for the vocabulary, and features are basically frequency information on each of those keywords. An alternative to overcome the problems of using machine learning is to take advantage of the semantic network of terms represented by Wikipedia's, exploring the articles relationships more deeply. We followed this direction in a second strategy (Section 3.3), which is built on top of a topic indexing method that uses Wikipedia for text analysis. The evaluation of such strategy is presented in the next section.

4.3 Results for the Topic Indexing Strategy

In order to validate the topic indexing strategy (Section 3.3), we present a performance evaluation of the process and its results. For the first test, we applied our geotagging method over the news collection considering only the list of Brazilian states as candidate places. Then we automatically checked if the result tag was the expected state. This approach was used only to observe the results presented in Section 4.3.1. Finally a manual evaluation using the best parameters from the first test was performed over a small subset of articles, but considering states and cities as resulting tags (Section 4.3.2). Basically we observe values of *acceptance* (percentage of documents tagged), *macro-f1* (harmonic mean of precision and recall, average of all documents) and *accuracy* (number of successfully tagged documents over total tagged documents) on the experiments.

4.3.1 Search Depth Influence on Performance

Our method has a feature that keeps it from tagging a document if no candidate places have been found in the graph search. We evaluated such behavior using the *acceptance* metric, which shows the percentage of the collection that could be tagged. One can

imagine that the deeper the graph search, the higher the acceptance rate. However, deeper searches cause the accuracy to drop. In an exploratory analysis, we verified that the Portuguese Wikipedia graph used in our experiments had 1,197,628 nodes, with a maximum depth of 61 levels. However, the average distance between any two nodes seems to be small. According to the authors of a tool called *Six degrees of Wikipedia*⁴, the average number of clicks to get from any English Wikipedia article to any other is 4.573. Also, by quickly analyzing a cumulative distribution of average nodes reached per level in the breadth-first search, we noticed that with a depth of only three levels, more than 50% of all nodes could be reached, describing a typical long tail distribution. Further studies need to be done on this issue. For the present work, we just consider these clues, knowing that there is a trade-off between maximum depth and precision. The deeper the graph exploration, the harder it gets to select places as geotags since we start to work with more connections, causing the connectivity information to be less discriminative for the tag selection.

In order to check the method's behavior with different depth levels, we experimented with the maximum depth set between 1 and 4, with no fine tuning of other parameters, and checked the *acceptance* in each one of them. As Table 4.5 shows, as the search goes deep, more documents are accepted by the geotagger. However, the *accuracy* decreases rapidly; this will be explored further next. Notice that with the maximum depth set to 1, there is no breadth-first search in the graph, only topic indexing. But since the topics have edges between them, this simple adjacency information from the first graph level is already useful to calculate the $Score_p$ of the places, which in this case will be found as topics in texts.

In depth levels greater than 1, we use the exponential decay parameter present on equation 3.8, so that deeper places have their $Score_p$ value diminished rapidly, according to how far they are from the document in the graph. In table 4.6 we considered a maximum depth of 2, and then we explored different values for the exponential decay. We can see that lower values lead to lower accuracy. The highest accuracy that we got was 54.68%, less than the accuracy found with maximum depth of 1. This suggests that the geographic evidence found deeper in the graph needs to be treated carefully, or it can interfere with the accuracy of the geotagging process.

As we can see, the exponential decay affects the performance in a positive way, since it changes the way $Score_p$ is calculated for places found deeper on the graph search. However, it seems this decay was not enough to handle both directly and indirectly related places and to achieve better accuracy. The best accuracy found with

⁴<http://www.netsoc.tcd.ie/~mu/wiki/>

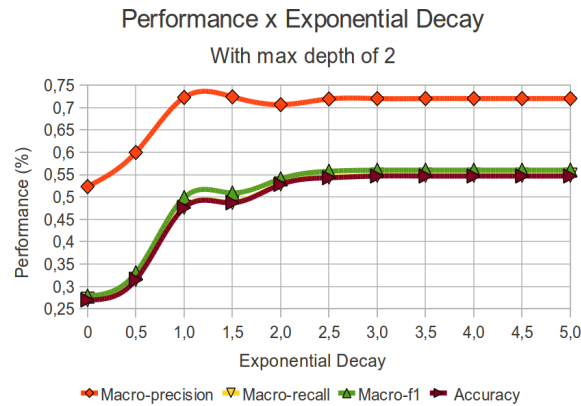


Figure 4.3. Geotagger performance for different exponential decay values for a *maxdepth* of 2

maximum depth 2 and varying the exponential decay was 54.68%, against 73.78% using one level of depth (i.e., no deep search in graph). As we said earlier, the number of nodes in the first few levels of depth grows quickly, and this can lead the tagger to errors. For instance, if an important city is found on the second level and it has numerous links, this adjacency can cause its $Score_p$ to overcome the score for less important cities found in the first level. To solve this, a more detailed study on how to normalize the graph adjacencies needs to be done, so that we can get a positive influence from deeper graph evidence in the geotagging process.

4.3.2 Manual Evaluation

In order to check the geotagger’s performance more closely, we evaluated three documents from each state to investigate if the given tags were adequately related to the text. For this evaluation, the set of target places was composed by all states and all cities from Brazil. Instead of the top-1 selection method used in the first experiments, we applied here the Relative to First Threshold, with top-3, and a threshold of 50%. In other words, the geotagger chose the best tag, and the next two ones ranked that had at least 50% of the the first tag’s $Score_p$. The geotagger was configured with a maximum depth of 1, not using the exponential decay, since it was the best tuning obtained from the previous tests. Documents rejected by the geotagger were not used, since we wanted to evaluate at least three documents per state.

After applying the geotagging procedure to all documents, we evaluated the tags document by document, reading its title and text, and analyzing the returned set of places. If any place was unrelated to the geographical scope of the news, we flagged

that document as an error. At the end, we got an accuracy of 78.57%. The work of Zong et al. (2005) is an important reference on associating places to texts using gazetteers. They achieved about 66% of accuracy with a collection of 50 documents with references to places of the the United States of America. Even though we do not assume their work as a baseline, their results suggests that the accuracy of our method is acceptable.

We analyzed each error case to consider their possible causes, and some interesting cases could be found, especially regarding the topic indexing disambiguation process. In one case, the tagger mistook the noun “*campanha*” (Portuguese for campaign) with a city also named *Campanha*. In another case, the city “Rio Branco” from *Acre* state, was found as part of a street name in another State, “Barão do Rio Branco” (name of the historical figure after whom both the city and the street were named). Another interesting mistake regards two states with a similar name: “Mato Grosso” and “Mato Grosso do Sul”, causing partial ambiguity and leading to many errors.

State	Website	Local Section Name	Collection Date	News count
Acre	http://www.agencia.ac.gov.br/	Municípios	July 2010	107
Alagoas	http://www.alagoasnoticias.com.br/	Municípios	July 2010	100
Amapá	http://www.amapadigital.net/	Geral	July 2010	97
Amazonas	http://www.noticiasdaamazonia.com.br/	Cidades	July 2010	100
Bahia	http://www.noticiasdabahia.com.br	Municípios	July 2010	98
Ceará	http://www.cearaagora.com.br/	Cidades, Interior	July 2010	100
Distrito Federal	http://www.correiobraziliense.com.br/	Cidades-DF	July 2010	100
Espírito Santo	http://www.sitebarra.com.br/	Geral	July 2010	100
Goiás	http://www.jornaldaimprensa.com.br/	Estado	July 2010	100
Maranhão	http://www.oimparcialonline.com.br/	Estado	July 2010	100
Mato Grosso	http://www.noticiando.com.br/	Municípios	July 2010	100
Mato Grosso do Sul	http://www.pantanalnews.com.br/	Cidades	July 2010	91
Minas Gerais	http://www.uai.com.br/	Minas	November 2009	104
Pará	http://www.paraonline.inf.br/	Notícias Pará	July 2010	100
Paraíba	http://www.paraiba1.com.br/	Cidades	July 2010	100
Paraná	http://www.parana-online.com.br/	Cidades	July 2010	89
Pernambuco	http://diariodepernambuco.com.br	Vida Urbana	July 2010	104
Piauí	http://piauinoticias.com/	Cidade	July 2010	100
Rio de Janeiro	http://g1.globo.com/rio	Geral	July 2010	100
Rio Grande do Norte	http://www.nominuto.com/	Cidades	July 2010	97
Rio Grande do Sul	http://www.diariodecanoas.com.br/	Cidades/Região	July 2010	100
Rondônia	http://www.rondoniagora.com/	Cidades	July 2010	100
Roraima	http://www.jota7.com/	Roraima	July 2010	100
Santa Catarina	http://www.folhanorte.com.br	All	July 2010	109
São Paulo	http://g1.globo.com/sao-paulo/	Geral	July 2010	102
Sergipe	http://emsergipe.globo.com/	Sergipe	July 2010	100
Tocantins	http://www.anoticia-to.com.br/	Cidades	July 2010	100
Total				2698

Table 4.1. Details about the test collection built with local news from each one of the 27 states from Brazil.

Technique	Different training set size (% of a fixed subset)				
	100%	80%	60%	40%	20%
TF-IDF	82.65	48.43	26.27	17.83	12.53
Wiki	84.1	81.45	77.83	70.6	60.48

Table 4.2. Accuracy for different training set sizes classifying the same test set

Random Seed Value	Number of Classes						
	2	3	4	5	6	7	8
1	97.09	84.69	88.56	89.69	88.69	89.29	88.81
2	96.12	85.34	88.56	88.52	89.5	89.97	88.69
3	94.66	84.36	88.32	88.52	88.85	89.7	87.73
4	95.15	83.06	88.83	89.3	90.47	90.11	88.45
5	94.66	83.71	88.81	88.91	88.69	90.11	88.33
Std. Dev.	1.05	0.88	0.37	0.51	0.76	0.35	0.42
Average	95.15	84.36	88.56	88.91	88.85	89.97	88.45

Table 4.3. TF-IDF strategy average accuracy considering a varying number of classes

Random Seed Value	Number of Classes						
	2	3	4	5	6	7	8
1	98.06	97.72	96.83	94.16	91.76	88.74	84.48
2	98.54	97.72	96.83	94.55	91.11	88.6	84.36
3	98.54	97.72	96.83	94.75	91.44	88.46	84
4	98.54	97.72	96.59	94.36	91.67	87.91	83.87
5	98.06	97.72	96.59	94.36	90.79	88.32	84.48
Std. Dev.	0.27	0	0.13	0.22	0.39	0.32	0.28
Average	98.54	97.72	96.83	94.36	91.44	88.46	84.36

Table 4.4. Wikipedia strategy average accuracy considering a varying number of classes

Max Depth	Acceptance	Accuracy
1	60,74%	73,78%
2	98,88%	49,87%
3	100,00%	11,11%
4	100,00%	4,81%

Table 4.5. Acceptance and accuracy for different *maxdepth* levels.

Exp. Decay	Macro-Precision	Macro-Recall	Macro-f1	Accuracy
0	52.33%	27.07%	27.82%	26.97%
0.5	59.96%	31.60%	33.09%	31.46%
1.0	72.26%	47.72%	49.88%	47.57%
1.5	72.40%	48.83%	50.91%	48.69%
2	70.66%	52.94%	54.07%	52.81%
2.5	71.92%	54.47%	55.76%	54.31%
3	72.01%	54.84%	56.00%	54.68%
3.5	72.01%	54.84%	56.00%	54.68%
4	72.01%	54.84%	56.00%	54.68%
4.5	72.01%	54.84%	56.00%	54.68%
5	72.01%	54.84%	56.00%	54.68%

Table 4.6. Geotagger performance for different exponential decay values for a *maxdepth* of 2

Chapter 5

Conclusions and Future Work

As a significant share of queries in Web searches present some geographic intention, it is important to conceive automatic ways to associate resources to places. We described two strategies to do so, using Wikipedia as a source of geographic evidence. A first strategy proposes an automatic text classification task based on the occurrence of keywords extracted from Wikipedia for a set of places. A second strategy intends to geotag texts with multiple place names by using a topic indexing technique based on Wikipedia articles. The topics are used to connect documents with the Wikipedia graph, allowing the search for related places. Evaluation experiments were presented over a collection of documents associated to Brazilian states, and we have demonstrated the feasibility of using Wikipedia as source of geographic evidence, taking advantage of free, up-to-date and wide knowledge.

Section 3.2 detailed the first strategy, and we presented a method for classifying text in a set of locations, based on evidence obtained from the titles of Wikipedia entries connected to place-related articles. We showed how this kind of geographic evidence can be used to build term lists that are used as classification features. Our approach does not require the use of gazetteers as sources of place names, and proposes the generation of term sets from a given set of places, which correspond to the classes in an automatic classification process. Experiments showed that a high level of precision can be achieved with this approach. Results with a relatively small collection showed that this method has very good potential, and also that there is room for much improvement. This work was detailed in a full paper (Alencar et al., 2010). The use of machine learning techniques presented some limitations, what confirms the conclusions of Silva et al. (2006), motivating us to work on a second strategy based on a method for topic indexing with Wikipedia.

Section 3.3 proposal presented a method for geotagging texts, based on the topic

indexing with Wikipedia method. By identifying Wikipedia articles as topics in the text, we connect the document to the encyclopedia's semantic network, and then we use a breadth-first search in the articles graph to obtain a list of candidate places related to the text. Then, we apply a scoring technique to determine the best places that should be given as returned tags. During this step, some documents can be rejected due to low geographic information. We think that this rejecting behavior is useful for Information Retrieval mechanisms, since when users ask for documents about a certain location, the search engine should be able to retrieve only documents tagged within some estimation of certainty, instead of forcing a classification approach that always associates a document to some place. In the experimental evaluation, we first explored the different levels of depth in the breadth-search step. We noticed that even though deep searches increase the acceptance of documents by providing more geographic evidence, they also decrease the global accuracy due to the diversity of places that start to be considered as candidates for tags. We decreased this confusing effect of deep searching by setting up an exponential decay which makes deep places (far from the text in the graph) less important than near ones. Nevertheless, the gain obtained by tuning this exponential decay parameter was not enough to make maximum depths greater than 1 useful. Everything indicates that popular places in low depths sometimes can mess up the metrics, as their adjacency counts are high enough to overlap unpopular but relevant places from the first levels. Thus, our best result was obtained using a max depth of 1 (i.e. only adjacent nodes were considered), getting 73.78% of accuracy and accepting 60.74% of the collection to tag. A more detailed manual evaluation was done with a small set of documents, using cities and states as candidates for multiple-tags results. In this case we could get 78.57% of accuracy and we identified some problems related to the disambiguation process of the topic indexing step. This work was detailed in a full paper (Alencar and Davis Jr, 2011).

Regarding the contributions of this dissertation, we reinforce, first, that our strategies for the geographic scope computation benefit from using free and updated knowledge created by the Wikipedia community. Also, our techniques innovate by making it possible to identify the connection between places and documents, even when the place names are not explicitly mentioned in the text. Our proposal can be faced as both an alternative and an extension to the use of gazetteers in the association of places to texts, since one can still use the classic place names recognition based on dictionaries, while Wikipedia can be used to add another perspective to the same problem. Another contribution is that the topic indexing with Wikipedia method proposed by Medelyan et al. (2008) was adapted for Portuguese texts successfully, what confirms

the advantages of using Wikipedia as a source for language-independent text analysis. Finally, the news collection we have built for testing our strategies, also contributes to advances on GIR research, since there is a lack of available experimenting sets in this community.

Since our experiments were mainly on news associated to Brazilian states, future work includes experimenting with intra-urban place, such as neighborhoods, landmarks and regions as target places. Also, the errors explored on the manual evaluation of topic indexing with Wikipedia strategy lead us to consider in future work a fine tuning of the topic indexing method used, or even the proposal of a biased topic indexing method focused on high disambiguation precision for place names.

Sets of terms and expressions considered as important keywords for places were used to feed the OntoGazetteer, an ontological gazetteer proposed by Machado et al. (2010). We intend to use this next-generation gazetteer to solve more sophisticated GIR challenges that require more than the simple recognition of place names in texts. Future work will focus on solutions based on both the classical place name recognition and semantic analysis using knowledge from alternative sources. Place-related terms from Wikipedia are part of our first attempts to have such knowledge stored in a gazetteer, so that it can be used by different applications who needs to perform GIR tasks.

In this dissertation we have focused on the identification of geographic evidence in texts, basically keywords and topic names related to places. But as mentioned in the Related Work (Chapter 2), other perspectives can be also explored when one tries to associate places to Web resources. Some of them are: URL domain analysis, IP address heuristics and parsing of address strings in Web pages content. In the future, we also intend to integrate many of these perspectives with our proposal, and then work on single tool to handle all these point of view of the geographic aspect of Web pages. This tool should also deal with the uncertainty around each one of these method's assertions, combining them to give a more accurate answer on geographic scope of documents on the Web.

References

- Ahlers, D. and Boll, S. (2008). Retrieving address-based locations from the web. In *Proceedings of the 2nd international workshop on geographic information retrieval, GIR '08*, pages 27--34, New York, NY, USA. ACM.
- Alencar, R. O. and Davis Jr, C. A. (2011). Geotagging aided by topic detection with wikipedia. In Geertman, S., Reinhardt, W., and Toppen, F., editors, *Advancing Geoinformation Science for a Changing World*, volume 1 of *Lecture Notes in Geoinformation and Cartography*, pages 461–477. Springer Berlin Heidelberg.
- Alencar, R. O., Davis Jr., C. A., and Gonçalves, M. A. (2010). Geographical classification of documents using evidence from wikipedia. In *Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10*, pages 12:1--12:8, New York, NY, USA. ACM.
- Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04*, pages 273--280, New York, NY, USA. ACM.
- Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In *Proceedings of the 17th international conference on world wide web, WWW '08*, pages 357--366, New York, NY, USA. ACM.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Blessing, A., Kuntz, R., and Schutze, H. (2007). Towards a context model driven german geo-tagging system. In *Proceedings of the 4th ACM workshop on geographical information retrieval, GIR '07*, pages 25--30, New York, NY, USA. ACM.
- Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., and Davis, Jr., C. A. (2007). Discovering geographic locations in web pages using urban addresses. In *Proceedings*

- of the 4th ACM workshop on geographical information retrieval, GIR '07*, pages 31--36, New York, NY, USA. ACM.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web, WWW7*, pages 107--117, Amsterdam, The Netherlands. Elsevier Science Publishers B. V.
- Buscaldi, D. (2009). Toponym ambiguity in geographical information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR '09*, pages 847--847, New York, NY, USA. ACM.
- Buscaldi, D. and Rosso, P. (2007). A comparison of methods for the automatic identification of locations in wikipedia. In *Proceedings of the 4th ACM workshop on geographical information retrieval, GIR '07*, pages 89--92, New York, NY, USA. ACM.
- Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., and Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *In proceedings of the ACM SIGMOD workshop on the Web and databases, WebDB '99*, pages 91--96.
- Cardoso, N., Silva, M. J., and Santos, D. (2008). Handling implicit geographic evidence for geographic ir. In *Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08*, pages 1383--1384, New York, NY, USA. ACM.
- Davis Jr, C. A. and Fonseca, F. T. (2007). Assessing the certainty of locations produced by an address geocoding system. *Geoinformatica*, 11:103--129.
- Delboni, T. M., Borges, K. A. V., Laender, A. H. F., and Davis Jr., C. A. (2007). Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, 11:377--397.
- Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In *Proceedings of the 26th international conference on very large data bases, VLDB '00*, pages 545--556, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. (2003). Categorizing web queries according to geographical locality. In *Proceedings of the twelfth international conference on information and knowledge management, CIKM '03*, pages 325--333, New York, NY, USA. ACM.

- Guillén, R. (2008). Geoparsing web queries. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 781--785, Berlin, Heidelberg. Springer-Verlag.
- Himmelstein, M. (2005). Local search: The internet is the yellow pages. *Computer*, 38(2):26--34.
- Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H. (2008). Modelling vague places with knowledge from the web. *International Journal of Geographic Information Science*, 22:1045--1065.
- Kasneji, G., Ramanath, M., Suchanek, F., and Weikum, G. (2009). The yago-naga approach to knowledge discovery. *SIGMOD Records*, 37:41--47.
- Leidner, J. L. (2007). Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41:124--126.
- Machado, I. M., Alencar, R. O., Campos Jr., R., and Davis Jr., C. A. (2010). An ontological gazetteer for geographic information retrieval. In *Proceedings of the XI brazilian symposium on geoinformatics, GeoINFO 2010*, Campos do Jordão, Brazil. SBC.
- McCallum, A. and Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *Learning for text categorization: papers from the 1998 AAAI workshop*, volume 752, pages 41--48.
- Medelyan, O., Witten, I. H., and D., M. (2008). Topic index with wikipedia. In *Proceedings of the AAAI workshop on Wikipedia and artificial intelligence, WIKIAI 2008*, Chicago, IL.
- Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on information and knowledge management, CIKM '07*, pages 233--242, New York, NY, USA. ACM.
- Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08*, pages 509--518, New York, NY, USA. ACM.
- Orengo, V. and Huyck, C. (2001). A stemming algorithm for the portuguese language. In *Proceedings of International Symposium on String Processing and Information Retrieval*, pages 180--186, Los Alamitos, CA, USA. IEEE Computer Society.

- Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.
- Sanderson, M. and Han, Y. (2007). Search words and geography. In *Proceedings of the 4th ACM workshop on geographical information retrieval*, GIR '07, pages 13--14, New York, NY, USA. ACM.
- Schockaert, S., De Cock, M., and Kerre, E. E. (2008). Location approximation for local search services using natural language hints. *International Journal on Geographic Information Science*, 22:315--336.
- Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *CEUS - Computers, Environment and Urban Systems*, 30:378--399.
- Smith, T. R. and Frew, J. (1995). Alexandria digital library. *Communications of ACM*, 38:61--62.
- Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005). Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on geographic information retrieval*, GIR '05, pages 17--24, New York, NY, USA. ACM.
- Witten, I. H. and Frank, E. (2000). *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Zhu, B., Ramsey, M., Chen, H., Hauck, R. V., Ng, T. D., and Schatz, B. (1999). Creating a large-scale digital library for geo-referenced information. In *Proceedings of the fourth ACM conference on digital libraries*, DL '99, pages 260--261, New York, NY, USA. ACM.
- Zong, W., Wu, D., Sun, A., Lim, E.-P., and Goh, D. H.-L. (2005). On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries*, JCDL '05, pages 354--362, New York, NY, USA. ACM.