Geographical classification of documents using evidence from Wikipedia

Rafael Odon de Alencar Clodoveu Augusto Davis Jr. Marcos André Gonçalves
Federal University of Minas Gerais
Av. Antônio Carlos, 6627
Belo Horizonte – MG - Brazil
odon.rafael@gmail.com, {clodoveu, mgoncalv}@dcc.ufmg.br

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 paper describes strategies for text classification into geography-related categories, using evidence extracted from Wikipedia. We use terms that correspond to entry titles and the connections between entries in Wikipedia's graph to establish a semantic network from which classification features are generated. Results of experiments using a news dataset, classified over Brazilian states, show that such terms constitute valid evidence for the geographical classification of documents, and demonstrate the potential of this technique for text classification.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: Miscellaneous; H.3.1 [Content Analysis and Indexing]: Indexing methods Classification Scheme: http://www.acm.org/class/1998/

General Terms
Algorithms, experimentation, performance.

Keywords
Geographic information retrieval. Text classification. Geospatial evidence.

1. INTRODUCTION
As the Internet continues to expand, the need for effective information retrieval becomes an even greater challenge. Search engines are constantly enhanced in order to respond to user queries with the best possible Web documents, taking into consideration various aspects about the keywords used in the query and their semantic meaning.

The need to obtain or approximate 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 portion of the users' demand [10, 21]. Wang et al. [24] state that many user activities on the Web are directly related to the user's location, thus it is important to conceive and to develop applications that take into consideration this intention. 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 [12, 22, 23]. Successfully accomplishing this task would give us means to enhance current indexing and retrieval mechanisms, so that people can search for documents that fall within a delimited geographic scope (i.e. perform *local search* [11, 22]), find nearby services or merchants, or 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.

This paper shows a technique for classifying documents according to their association to places, based on the occurrence of terms that coincide with Wikipedia entry titles. This technique does not employ the entire text from Wikipedia, but uses titles and hypertext connections between entries, a.k.a. *Wikipedia's graph*. We demonstrate the feasibility of the technique and the potential for precise results through an experiment, which classifies news articles according to the Brazilian states to which they refer, and compare the results to a bag-of-words approach.

This paper is organized as follows. Section 2 shows related work. Section 3 presents the proposed technique, while Section 4 presents experiments and their results. Section 5 presents some conclusions and describes future work.

2. RELATED WORK
Wang et al. [24] consider three different types of evidence to determine the geographic location(s) associated with a 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 GeoIP[1]. Second, the location can be inferred by analyzing the textual content of the document. Third, 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. We are particularly interested in the second type, since the location of the Web server can be completely unrelated to the subject of the document, and due to the fact that IP locating techniques are sometimes error-prone and imprecise.

GIR'10, 18-19th Feb. 2010, Zurich, Switzerland.

[1] http://www.maxmind.com/app/ip-location

Borges et al. [4] 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 the recognition of 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 [1, 3, 26], then performing some sort of geocoding [9].

Silva et al. [23] 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 [5] 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 Names[2].

Beyond the recognition of place names, we observe that many other terms that can occur in 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 between places that share the same name. The feasibility of this idea was explored by Backstrom et al. [2], 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. [8] call these terms *implicit geographic evidence*.

Recent work has shown that Wikipedia[3] can be a valuable source of information, considering the semi-structured annotations that exist in its entries and the usual richness of outgoing links, which compose a *de facto* semantic network. Kasneci et al. [13] 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 WordNet[4]. Milne and Witten [16] present a text enrichment strategy that automatically adds hyperlinks to Wikipedia entries, in a process known as *wikification* [15]. This is done without any parsing or natural language analysis, only matching terms from the text to Wikipedia entries. Cardoso et al. [8] 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".

Buscaldi et al. [7] use the Wikipedia as a rich source of geographical evidence and they propose to build a geographical ontology considering geo-political entities found in the encyclopedia's entries. In another work, Buscaldi and Rosso [6] compare different methods to automatic identify geographical articles in Wikipedia. Our work considers that the Wikipedia graph can generate many semantic connections useful to gather textual evidence for places. We explore connections that start from or are directed to entries that refer to places, and demonstrate a process to extract lists of terms that can be used to recognize references to that place. Of course, some of these terms are common to many entries, while others can be very specific. We propose a way to determine how specific terms are in respect to a place, and use this information to classify documents according to their likely relationship to places. The method is presented in the next section.

3. TEXT CLASSIFICATION USING EVIDENCE FROM WIKIPEDIA

A method for the identification of the geographic context of a document is expected to indicate the place (or places) more likely to be referenced by the document, even if the place name does not appear explicitly in the document. The technique we present here sets itself apart from others in the literature because it does not involve the recognition of direct references to places in the text, using a gazetteer as a source of place names. It tries to find location indications from the occurrence of terms related to a place through a Wikipedia entry. Since we use a single-label classification approach, results from the proposed method indicate the best match between the document and one of a predetermined set of places (classes), which may or may not have been obtained from a gazetteer. The set of places is problem-dependent. For instance, in the case study presented in this paper, we used the set of 27 Brazilian states.

First, we gather place-related terms from Wikipedia, by exploring the graph formed by links to and from entries that refer to the each place of interest. For each hyperlink pointing to or pointed by the entry about a place, the related entry's title is kept as a term related to the place. We present a strategy for generating classification features from the terms, associating each term with an estimation of its discriminative power, which varies because many of the terms can be related to more than one place. Of course, terms that are related to fewer places should weigh more in the classification scheme, but other terms can be important too, especially for disambiguation. The intention of this proposal is not to create a definitive geographic classifier for texts, but rather to demonstrate the validity of using evidence obtained from the Wikipedia, seen as a semantic network, and to generate further questions and ideas for improvement.

More formally, the problem involves the classification of a set of n documents $D = \{d_1, d_2, \ldots, d_n\}$ as referring to one of a set of m places $P = \{p_1, p_2, \ldots, p_m\}$. For each place p_i, we find a corresponding Wikipedia entry, and obtain from the linkage graph two sets of terms, which correspond to the titles of entries which reference the place's entry or are referenced by it. We call these the *inlink* and the *outlink* term sets. The generation of the term sets is explained in the next section.

[2] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[3] http://www.wikipedia.org

[4] http://wordnet.princeton.edu

3.1 Geographic Evidence from Wikipedia

The Wikipedia can be seen as a large graph, in which each node n_i represents an entry, and each edge e_{ij} represents a hyperlink connecting node i to node j. Since each entry is created and maintained by the user community, we assume that the links are also the subject of scrutiny, and therefore we consider the graph structure itself to be a rich network of semantically related terms.

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 entries themselves, including attributes such as an identifier, title, and namespace (type). The *pagelinks* table stores the connections between entries. We used the titles of entries that are adjacent to an entry 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.

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. 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 a reference, 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 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 it to the list of terms if it is found in a link from the Wikipedia graph.

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

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 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 [14]. 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 [14].

The idea was 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 sets of features. Equations 2 and 3 are used to determine the value of each single feature corresponding to a given place/document pair.

$$S^{in}(p_i, d_j) = \sum_{l=1}^{|in|} w(t_l) \times \text{Frequency}(t_l, d_j) \qquad (2)$$

$$S^{out}(p_i, d_j) = \sum_{l=1}^{|out|} w(t_l) \times \text{Frequency}(t_l, d_j) \qquad (3)$$

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)* will add significantly to the sum, while terms with low weights can also contribute to the result, but with a smaller impact. We would 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 our motivation to use an automatic classifier came from the need to fine tune a frequency-based information model, we decided to improve the proposed classification model to 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 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 1). The inlinks and outlinks are obtained from the text of the entries, and then weights are calculated according to Equation 1 (Figure 2).

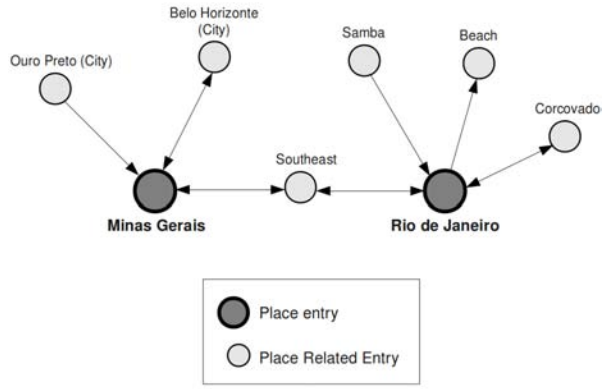

Figure 1 - Wikipedia graph example

For the classification, consider the documents shown in Figure 3. 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.

Minas Gerais inlinks	
Article	w(t)
Ouro Preto (city)	1
Belo Horizonte (city)	1
Southeast	0.25

Rio de Janeiro inlinks	
Article	w(t)
Samba	1
Corcovado	1
Southeast	0.25

Minas Gerais outlinks	
Article	w(t)
Belo Horizonte (city)	1
Southeast	0.25

Rio de Janeiro outlinks	
Article	w(t)
Beach	1
Corcovado	1
Southeast	0.25

Figure 2 -Term weights in the inlinks and outlinks sets for each place

Document 1

Ouro Preto and **Belo Horizonte** are located in Minas Gerais.

Document 2

Samba music is very popular in Rio de Janeiro, a Brazilian **southeast** state.

Figure 3 - Sample documents for classification

The first classification model would then generate the features for the classifier, as shown in Table 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 classifier would use 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.

Table 1 – Features in the first strategy, and true classes for each document

Document 1	Minas Gerais		Rio de Janeiro		Class
	in	out	in	out	
Document 1	2	1	0	0	Minas Gerais
Document 2	0.25	0.25	1.25	0.25	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 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 2 and 3, but this time the summation is performed over groups of terms that have the same discriminative power.

Table 2 – Features in the second strategy and classification results

Document 1	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

In a preliminary test of the first classification model, performed on a set of 200 documents and two places, the Multinomial Naïve Bayes classifier was able to accurately classify 92.7% of the documents (average precision in a 10-fold cross validation [25] set of tests). With the second and more flexible classification model, the Multinomial Naïve Bayes classifier reached an average precision rate of 98.0% with 10-fold cross validation, over the same collection used in the previous test. The improved precision shows that having more evidence to support the classification can lead to better results, as expected. However, the growth of the number of features needs to be taken into consideration.

Next section shows a larger experiment, involving a larger collection and a larger number of classes, and a comparison to a text classification that uses no geographic evidence as a baseline.

4. EXPERIMENTAL EVALUATION

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, as described in Section 4.1. We then performed an automatic classification using a bag-of-words approach with a TF-IDF weighting scheme to serve as a baseline (Section 4.2). Finally, we applied the proposed method to the same documents and compared the classification results to the baseline (Section 4.3).

4.1 Creating the document collection

In order to create a document collection for experimentation, we selected newspaper sites from various Brazilian states, and retrieved from each of them a number of articles from their "local and state news" sections (Table 3). We assumed that the source of the news document indicates its true class directly; for instance, the *Minas* section of the news site *www.uai.com.br* is expected to contain only news about Minas Gerais state. Crawlers were prepared for each news source, so that custom regular expressions were used to extract the contents of news pages. We only extracted the text from the main portion of the news item, i.e., its title, author (if available), publication date, content paragraphs and possibly other elements embedded in the text pane. Less interesting and repetitive portions of each Web Page, such as navigation menus, heading, footers, and advertising were eliminated.

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 key words. We used the stopword list for Portuguese compiled by the Snowball project [19]. 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 [18], as implemented in the PTStemmer project [17]. The same treatment was applied to the titles of entries from Wikipedia, in order to allow the matching of terms in the documents. Table 3 lists the sources of the 831 news documents we included in the collection.

Although the size of the collection may be considered small, we tried to ensure that the collected documents were geographically related, so that the classification success can be adequately verified. The existence of geographic references was achieved first by carefully choosing the sources for collection, and then by performing a simple visual inspection of the articles title to ensure that there were not many general subject news. Considering the magnitude of the effort to create the dataset and the time available for experimental tests, we collected news from only 8 of the 27 Brazilian states, including the three more important ones (Minas Gerais, Rio de Janeiro and São Paulo), and others which are representative of the other geopolitical regions of the country. In future work, we intend to consider the full set of 27 states, and to experiment with other spatial granularities (e.g. cities, regions). Such experimentation will require the creation of a larger and more detailed set of geographically-labeled documents

4.2 Baseline: traditional 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 [20]. 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 3). 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). The final measurement is the product of the two components for each document (Equation 5). TF-IDF results are low for common terms, and higher for rare terms.

Table 3 – Sources of the document collection

Brazilian State	News Web Site	Local News Section Name	Access Date	# Docs
Minas Gerais (MG)	www.uai.com.br	Minas	09/11/09	104
São Paulo (SP)	g1.globo.com	São Paulo	09/11/09	101
Rio de Janeiro (RJ)	g1.globo.com	Rio	09/11/09	102
Acre (AC)	www.ac24horas.com.br	Acre	10/11/09	103
Amazonas (AM)	www.amazonasagora.com.br	Municipios	25/11/09	103
Bahia (BA)	ibahia.globo.com	Bahia	30/11/09	105
Santa Catarina (SC)	www.folhanorte.com.br	All	01/12/09	109
Pernambuco (PE)	www.diariodepernambuco.com.br	Vida Urbana	01/12/09	104
			Total	831

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (3)$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \square d\}|} \qquad (4)$$

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \qquad (5)$$

Several classifiers could be used with this document representation. We chose Multinomial Naive Bayes Classifier due its popularity, simplicity, and ease of operation to obtain good results when working with term frequencies [14]. 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 success rate 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 [14]. This result is used as a baseline for tests using our method, as shown in the next section.

4.3 Experiments and results

Experiments were performed using Weka [25] 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 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 intention 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 to be 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 to the TF-IDF model presented in section 4.2.

First, we performed a test on the capacity for each model to succeed with varying training set sizes. We divided the 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 test allowed us to verify how much training data is necessary for each classification strategy to achieve successful results. Figure 4 and Table 4 show the results of this test.

Table 4 - Success rate for different training set sizes classifying the same test set

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

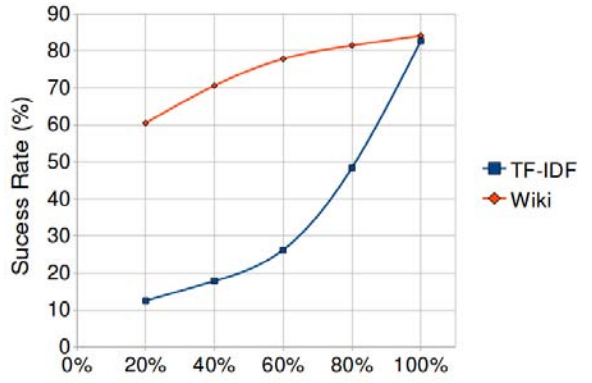

Figure 4. Success rate versus training set sizes

Notice that the classification strategy using Wikipedia evidence starts with a success rate 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 success rate only approaches the one obtained using Wikipedia evidence when all of the data are used for training. 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 success rate 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 5, Table 5, and Table 6 show the results.

Table 5 - TF-IDF strategy average success rate considering a varying number of classes

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	87.83	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 6 - Wikipedia strategy average success rate 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	91.6	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

Notice that the classifier based on Wikipedia evidence showed good results for few classes, but its performance drops as new classes are added. The decrease 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 curve shows a lower region 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 3). 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 success rate 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.

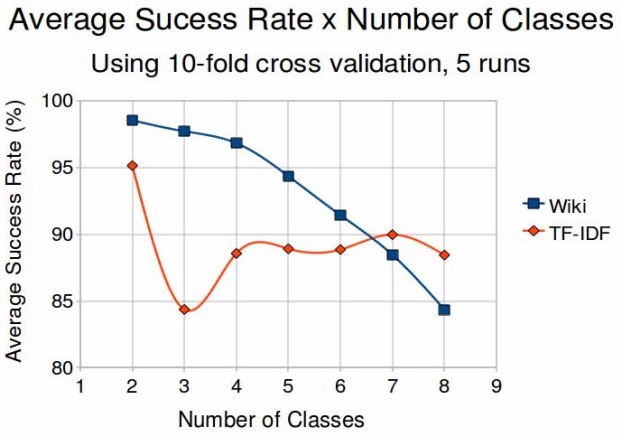

Figure 5 – Average success rate vs. number of classes

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 some 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 the geographical 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 the 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, and again 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 success rate of 82.4%, i.e., about 6% less than the previous result. The Wikipedia-based method obtained a success rate 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 will cause the TF-IDF method's success rate 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, a small test collection was created using the text of the Wikipedia entries for the each one of the 8 Brazilian states. The training set was composed of the initial 831 news documents. The Wikipedia entries 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% success rate, 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 entry. However, this test shows that traditional text classification is clearly outperformed by the method proposed in this paper. While our method gets better results in the presence of geographical evidence, the traditional method tends to improve in the presence of non-geographical evidence, which may not exist in documents that significantly differ from the ones used in training.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a method for classifying text according to geographic location, based on evidence obtained on the connections between Wikipedia entries and their titles. 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 the 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 a very good potential, and also that there is room for much improvement. Future work includes the generation of a much larger collection, using a wider variety of sources, so that common characteristics of documents from the same source, such as elements of structure and vocabulary, cannot interfere with the classification. It is also interesting to evaluate the possibility of combining both classification strategies (TF-IDF of a bag-of-words and Wikipedia evidence).

The strategy for extracting geographic evidence from Wikipedia can be improved in several ways. One of them would be to expand the list of relevant terms by looking up alternative terms for the same entity, which exist in redirecting and disambiguation resources. Another idea to be explored is to verify the terms used to create the hypertext links between Wikipedia entries, since they

can be different from the entry's title. Furthermore, the semantic network of terms represented by Wikipedia's graph can be explored more deeply. Some of these improvements may have a positive impact on the drop of precision we got in the first experiment, since the impact of terms that are adjacent to two or more places will probably be reduced. This work considered only adjacent entries and their relative importance, but other techniques and metrics related to complex networks can also be used.

6. ACKNOWLEDGMENTS

The authors acknowledge the support from FAPEMIG, CNPq and CAPES, Brazilian agencies in charge of fostering research and development, in the form of individual research grants and scholarships, and also the support from the Brazilian National Institute of Science and Technology for the Web (CNPq grant 573871/2008-6).

7. REFERENCES

[1] Ahlers, D. and Boll, S. *Retrieving address-based locations from the web*. in *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*. 2008. Napa Valley, CA, USA.

[2] Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. *Spatial Variation in Search Engine Queries*. in *International World Wide Web Conference (WWW)*. 2008. Beijing, China.

[3] Blessing, A., Kuntz, R., and Schütze, H. *Towards a context model driven German geo-tagging system*. in *Proceedings of the 4th ACM Workshop in Geographical Information Retrieval*. 2007. Lisbon, Portugal.

[4] Borges, K.A.V., Laender, A.H.F., Medeiros, C.B., and Davis, C.A. *Discovering Geographic Locations in Web Pages Using Urban Addresses*. in *Proceedings of the 4th ACM Workshop on Geographic Information Retrieval*. 2007. Lisbon, Portugal.

[5] Brin, S. and Page, L. *The anatomy of a large hypertextual Web search engine*. in *Proceedings of the 7th International Conference on the World Wide Web*. 1998. Brisbane, Australia.

[6] Buscaldi, D. and Rosso, P. *A Comparison of Methods for the Automatic Identification of Locations in Wikipedia*. in *Proceedings of the GIR 2007 Workshop*. 2007. Lisbon, Portugal.

[7] Buscaldi, D., Rosso, P., and Peris, P. *Inferring Geographical Ontologies from Multiple Resources for Geographical Information Retrieval*. in *Proceedings of the 3rd GIR Workshop, SIGIR'06*. 2006. Seattle, WA, USA.

[8] Cardoso, N., Silva, M.J., and Santos, D. *Handling implicit geographic evidence for geographic information retrieval*. in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*. 2008. Napa Valley, CA, USA.

[9] Davis Jr., C.A. and Fonseca, F.T., *Assessing the Certainty of Locations Produced by an Address Geocoding System.* Geoinformatica, 2007. **11**(1): p. 103-129.

[10] Delboni, T.M., Borges, K.A.V., Laender, A.H.F., and Davis Jr., C.A., *Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions.* Transactions in GIS, 2007. **11**(3): p. 377-397.

[11] Himmelstein, H., *Local Search: The Internet is the Yellow Pages.* IEEE Computer, 2005. **38**(2): p. 26-35.

[12] Jones, C.B., Purves, R.S., Clough, P.D., and Joho, H., *Modelling vague places with knowledge from the Web.* International Journal of Geographical Information Science, 2008. **22**(10): p. 1045-1065.

[13] Kasneci, G., Ramanath, M., Suchanek, F., and Weikum, G., *The yago-naga approach to knowledge discovery.* SIGMOD Record, 2008. **37**(4): p. 41-47.

[14] McCallum, A. and Nigham, K. *A comparison of event models for naive Bayes text classification*. in *AAAI-98 Workshop on Learning for Text Categorization*. 1998: Tech. Rep. WS-98-05, AAAI Press.

[15] Mihalcea, R. and Csomai, A. *Wikify!: linking documents to encyclopedic knowledge*. in *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*. 2007. Lisbon, Portugal.

[16] Milne, D. and Witten, I.H. *Learning to link with Wikipedia*. in *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'08)*. 2008. Napa Valley, CA, USA.

[17] Oliveira, P. *PTStemmer*. 2009 [cited Dec 3 2009]; Available from: http://code.google.com/p/ptstemmer/.

[18] Orengo, V.M. and Huyck, C. *A stemming algorithm for the Portuguese language*. in *String Processing and Information Retrieval (SPIRE 2001)*. 2001.

[19] Porter, M. *Snowball*. 2005 [cited Dec 3 2009]; Available from: http://snowball.tartarus.org/index.php.

[20] Salton, G. and McGill, M.J., *Introduction to Modern Information Retrieval*. 1983, New York, NY: McGraw Hill Book Co.

[21] Sanderson, M. and Han, Y. *Search words and geography*. in *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval (GIR'07)*. 2007. Lisbon, Portugal.

[22] Schockaert, S., De Cock, M., and Kerre, E.E., *Location approximation for local search services using natural language hints.* International Journal of Geographic Information Science, 2008. **22**(3): p. 315-336.

[23] Silva, M.J., Martins, B., Chaves, M., Cardoso, N., and Afonso, A.P., *Adding Geographic Scopes to Web Resources.* Computers, Environment and Urban Syst., 2006. **30**: p. 378-399.

[24] Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W. *Detecting Geographic Locations from Web Resources*. in *Proc. of the 2nd Int'l Workshop on Geographic Information Retrieval*. 2005. Bremen, Germany.

[25] Witten, I.H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2nd ed. 2005: Morgan Kaufmann.

[26] Zong, W., Wu, D., Sun, A., Lim, E., and Goh, D.H.G. *On Assigning Place Names to Geographic Related Web Pages*. in *Proc. of the 5th ACM/IEEE-CS Joint Conf. on Digital Libraries*. 2005. Denver, Colorado, USA.