Classifying Documents According to Locational Relevance

Ivo Anastácio, Bruno Martins, and Pável Calado

Instituto Superior Técnico, INESC-ID,
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal
{ivo.anastacio,bruno.g.martins,pavel.calado}@ist.utl.pt

Abstract. This paper presents an approach for categorizing documents according to their implicit locational relevance. We report a thorough evaluation of several classifiers designed for this task, built by using support vector machines with multiple alternatives for feature vectors. Experimental results show that using feature vectors that combine document terms and URL n-grams, with simple features related to the locality of the document (e.g. total count of place references) leads to high accuracy values. The paper also discusses how the proposed categorization approach can be used to help improve tasks such as document retrieval or online contextual advertisement.

Key words: Document Classification, Geographic Text Mining

1 Introduction

Automated document classification is a well studied problem, with many applications in text mining and information retrieval [11]. A recent trend in text mining applications relates to extracting geographic context information from documents. It has been noted that the combination of techniques from text mining and geographic information systems can provide the means to integrate geographic data and services, such as topographic maps and street directories, with the implicit geographic information available in Web documents [2,6,9].

In this work, we propose that textual documents can be characterized according to their implicit *locational relevance*. For example, a document on the subject of computer programing can be considered *global*, as it is likely to be of interest to a geographically broad audience. In contrast, a document listing pharmacies or take-away restaurants in a specific city can be regarded as a *local*, i.e., likely to be of interest only to an audience in a relatively narrow region. Somewhere in between is a document describing touristic attractions in a specific city, likely to be of interest to both the inhabitants of that city and to potential visitors from other parts of the world. In the context of this work, locational relevance is, therefore, a score that reflects the probability of a given document being either

This work was partially supported by the FCT (Portugal), through project grant PTDC/EIA/73614/2006 (GREASE-II).

global (i.e., users interested in the document are likely to have broad geographic interests) or local (i.e., users interested in the document are likely to have a single narrow geographic interest). This score can be produced from the confidence estimates assigned by a binary classifier such as a Support Vector Machine [12].

Previous research has addressed the problem of automatically computing geographic scopes of Web documents [1,2]. Techniques have also been proposed for detecting locationaly relevant search engine queries [3,4]. However, to the best of our knowledge, no description has ever been published on techniques for classifying documents according to locational relevance (i.e., classifying documents as either local or global). This is a significantly different problem from that of assigning documents to geographic scopes, since two documents can have the same scope but different locational relevances. For instance, the Web page of a research group in Lisbon and the Web page of a local restaurant in Lisbon have the same geographic scope, nonetheless, people visiting the restaurant's page are most probably taking into consideration the location, while people visiting the researcher's page are most probably interested in their studies, regardless from where the group is physically located.

To solve this problem, we propose an approach for categorizing documents according to their implicit locational relevance, using state-of-the-art machine learning techniques. We report a thorough evaluation of several classifiers, built using support vector machines, and explore many alternative features for representing documents. In addition, we also discuss how our classifier can be used to help improve tasks such as document retrieval or online advertisement.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes our classification approach, detailing the proposed features. Section 4 presents and discusses the experimental validation, also describing applications for locational relevance classifiers. Finally, Section 5 presents our conclusions and directions for future work.

2 Related Work

Traditional Information Retrieval and Machine Learning research has studied how to classify documents according to predefined categories [10,11]. The subarea of Geographic Information Retrieval has addressed issues related to the exploitation of geographic context information mined from textual documents. In this section, we survey relevant past research on these topics.

2.1 Document Classification

Document classification is the task of assigning documents to topic classes, on the basis of whether or not they share some features. This is one of the main problems studied in fields such as text mining, information retrieval, or machine learning, with many approaches described in the literature [10,11]. Some methods suitable for learning document classifiers include decision trees [14], logistic regression [15]

and support vector machines [7,12]. SVMs can be considered a state-of-the-art method in binary classification, returning a confidence score in the assigned class.

Previous works have also suggested that in text domains, due to the high-dimensionality of the feature space, effective feature selection can be used to make the learning task more efficient and accurate. Empirical comparisons of different feature selection methods have been made in the past [13,16], with the results suggesting that either Chi-square or information gain statistics can provide good results. In this work, we use feature selection methods in order to examine the most discriminative features.

2.2 Mining geographical information from text documents

Previous research in the area of geographic information retrieval has addressed problems such as the recognition and disambiguation of place references present in text, and the assignment of documents to encompassing geographic scopes.

Leidner presented a variety of approaches for handling place references on textual documents [9]. The problem is usually seen as an extension of the named entity recognition (NER) task, as proposed by the natural language processing community [17,18]. More than recognizing mentions to places in text, which is the subject of NER, the task also requires for the place references to be disambiguated into the corresponding locations on the surface of the Earth, i.e., assigning geospatial coordinates to the place references [9]. Place reference disambiguation usually relies on gazetteer matching, together with heuristics such as default senses (i.e., disambiguation should be made to the most important referent, based on population counts) or spatial minimality (i.e., disambiguation should minimize the convex hull that contains all candidate referents) [9,19]. Metacarta[1] is a commercial company that sells state-of-the-art geographic information retrieval technology. The company also provides a freely-available *geotagger* Web service that can be used to recognize and disambiguate place references in text. An early version of the Metacarta *geotagger* has been described by Rauch et al. [20]. Yahoo! Placemaker[2] is another free Web service which provides recognition and disambiguation of place references in text. The complementar Yahoo! GeoPlanet[3] Web service is an example of an online gazetteer, returning descriptions of places based on their name.

Anastácio et al. surveyed different approaches for assigning documents to geographic scopes [28]. In one of the pioneering works in the area of geographic information retrieval, Woodruff and Plaunt proposed a technique with basis on the place references discovered in the text [6]. Their method was based on disambiguating the place references into the bounding polygons that correspond to the geographic area of the referents. The geographic scope of the document is afterward computed by overlapping the areas of all the polygons. More recently, Ding et al. proposed specific techniques for extracting the geographical scope of web

[1] `http://ondemand.metacarta.com`
[2] `http://developer.yahoo.com/geo/placemaker`
[3] `http://developer.yahoo.com/geo/geoplanet`

pages [1]. For example, the *Diário de Coimbra* online newspaper has a geographical scope that consists of the city of Coimbra, while the *Publico* newspaper has a geographical scope that includes the entire territory of Portugal. To compute the geographical scope of a web document, Ding et al. propose two complementary strategies: (1) a technique based on the geographical distribution of HTML links to the page, and (2) a technique based on the distribution of geographical references in the text of the page. Amitay et al. also proposed a technique for assigning Web documents to the corresponding geographic scope [2], leveraging on *part-of* relations among the recognized place references (i.e. *Lisbon* and *Porto* are both part of *Portugal*, and documents referring to both these places should have *Portugal* as the scope). Looping over the disambiguated references, this approach aggregates, for each document, the importance of the various levels in a location hierarchy. The hierarchy levels are then sorted by importance and results above a given threshold are returned as the geographic scope.

Gravano et al. proposed a technique for classifying search engine queries as either local or global, using the distributional characteristics of location names occurring in the results produced by a search engine to the query [3]. There are many similarities between the work by Gravano et al. and the proposal of this paper, but here we are instead concerned with classifying documents as global or local, instead of classifying user queries.

3 Classifying documents according to locational relevance

Assigning documents to global and local classes, according to their implicit locational relevance, is a hard document classification problem. Instead of just applying a standard classification approach, based on a bag-of-words representation of the documents, we argue that specific geographic features are also well suited to reflect the locational characteristics of the documents.

Global documents often do not include any mentions to place names. Consider the home page of the Weka software package[4]. Users reading this document are probably looking for tutorials about machine learning, and they are not restricted in their interests to a specific geographic scope. Nevertheless, it is interesting to note that global documents can sometimes include mentions to place names. Consider a document describing a review of U2's latest concert in Lisbon. The location name is clearly distinguishable in the document, but the readers may have completely different geographic interests.

Local documents are, on the other hand, more likely to contain mentions to place names, particularly place names associated to small regions. Local documents are also more likely to contain references to places that are restricted to a somewhat confined area, whereas global documents can contain place references to distinct places around the world. Examples of local documents include local business listings or descriptions of local events.

[4] `http://www.cs.waikato.ac.nz/ml/weka`

3.1 Classification Features

The feature vectors used in the proposed classification scheme combine information directly extracted from the full text of the documents, or from the document URL, with higher level geographic information, mined from the documents using techniques from the area of geographic information retrieval. We group the considered features in four classes, namely (1) textual features, (2) URL features, (3) simple locative features, and (4) high level locative features. Textual and URL features are directly extracted from either the text or the URL for the document, whereas the remaining require geographic text mining.

In the case of the textual features, the idea was to capture the thematic aspects, encoded in the document's terminology, that can influence the decision of assigning a document to either a global or local class. For instance documents about restaurants or pharmacies are more likely to be local than documents about programming languages or music downloads.

The Yahoo! Term Extraction[5] Web service, a state-of-the-art industrial tool for key term extraction, was used to discover important words in the documents. Its implementation is available via an open Web service, which takes a text document as input and returns a list of significant words or phrases extracted from the document.

The full set of textual features is shown below:

- Word stems occurring in the lowercased document text, weighted according to the term frequency vs. inverse document frequency scheme (TF/IDF). Stopwords were removed according to the list provided by the Weka package.
- Lowercased words selected by the Yahoo! Term Extraction service as the most important in the document, weighted through the TF/IDF scheme.

When classifying Web documents, another source of information that can be used for classification is their Uniform Resource Locator (URL). Previous research has shown that classifiers built from features based solely on document URLs can achieve surprisingly good results on tasks such as language identification [24] or topic attribution [23]. Intuitively, URLs contain information that can be used to discriminate between local and global pages, such as top level domains or words such as *local* or *regional*. For instance, a document whose URL has a top level domain `.uk` is more likely to be local than a document with a top level domain such as `.com`. Taking inspiration on the experiments reported by Baykan et al. [23], the following features were considered:

- Character n-grams, with n varying between 4 and 8, extracted from the lower-cased document URLs and weighted according to the TF/IDF scheme.

Simple locative features essentially correspond to counts for locations recognized in the documents, through the use of the geographic text mining services provided by Yahoo!. The Placemaker text mining service provides functionalities

[5] `http://developer.yahoo.com/search/content`

for recognizing and disambiguating place references in text, returning the latitude and longitude coordinates for each place recognized. Using the GeoPlanet gazetteer service, we can expand the information returned by Placemaker with elements such as the county, state, country, continent and bounding polygon corresponding to each of the recognized place references. The combined functionalities of these two services allow us to experiment with a wide variety of locative document features.

It is important to notice that using the Yahoo! Placemaker for recognizing the place references over text has some advantages compared to the usage of a simpler dictionary-based approach. Since Placemaker uses natural language contextual clues, it helps to disambiguate whether a word like *reading* refers to the location in England or to the verb sense of *to read*. Placemaker's *geotagger* also covers many colloquial place names (e.g. *nyc* for *New York City*), as well as interest points (e.g. *Eiffel Tower*) that may appear in the text.

Having locations referenced in the document can indicate a tendency towards a higher locality, particularly if these locations are all related to a single relatively narrow region. On the other hand, having no locations whatsoever, or having many locations from different parts of the World, can indicate that the document has a global scope. We combine the location counts in various ways, aggregating the places according to containment relationships in order to group together the information for places in the same administrative divisions. We count the frequency of place references at different levels of detail (i.e., continent, country, state, city), as well as the aggregate total for all different unambiguous place references. The complete set of considered features is as follows:

- Total number of locations referenced in the text.
- Total number of unique locations referenced in the text.
- Number of unique locations, grouped by city, county, state, country and continent.
- Number of locations, grouped by city, county, state, country and continent.
- Number of unique locations, grouped by city, county, state, country and continent, considering the aggregated sub-locations that are hierarchically below (i.e., the number of counties includes the number of cities referenced in the text, the number of states includes the number of counties plus the number of cities, and so on).
- Total number of locations, grouped by city, county, state, country and continent, considering the aggregated sub-locations that are hierarchically below.

High level locative features correspond to the encompassing geographic areas computed with basis on the place references that were recognized in the text. The idea is that documents having broad geographic areas are more likely to correspond to global pages. Yu and Cai presented similar ideas for measuring the importance of geographic references in search engine queries [26]. The considered features are as follows:

- Area for the geospatial region corresponding to the geographic scope of the document, computed with the method proposed by Amitay et al. in the context of the Web-a-Where system [2].

- Area for the geospatial region corresponding to the geographic scope of the document, computed with the method proposed by Woodruff and Plaunt in the context of the GIPSY system [6].
- Area for the geospatial region that covers all the place references extracted from the document.
- Confidence score assigned by the Web-a-Where algorithm to the geographic scope that was computed for the document.
- Confidence score assigned by the GIPSY algorithm to the geographic scope that was computed for the document.

As stated in the description of the high level locative features, we implemented the scope assignment approaches proposed by Amitay et al. and by Woodruff and Plaunt [2,6]. The required geospatial computations were implemented through the use of the Java Topology Suite, an API of 2D spatial functions that supports the computation of area aggregates and intersections [8].

3.2 The Classification Method

In this work, we use Support Vector Machines (SVMs) for document classification. SVMs have been found quite effective for this task, which is characterized by having a high dimensionality in terms of the feature vectors [12]. SVM classifiers conceptually convert the original measurements of the features in the data to points in a higher-dimensional space that facilitates the separation between two classes. While the transformation between the original and the high-dimensional space may be complex, they need not to be carried out explicitly. Instead, it is sufficient to calculate a kernel function that only involves dot products between the data points, transformed from the original feature space. Commonly used functions include linear or Gaussian (radial basis) kernels, with the latter being recommended for document classification problems [12]. Determining the optimal classifier is equivalent to determining the hyper-plane that maximizes the total distance between itself and representative transformed data points (i.e., the support vectors). In our experiments, we used the Weka SVM implementation [5] with a Gaussian kernel function.

4 Experimental Evaluation

To evaluate our proposal, a large set of experiments was performed. This section describes the reference datasets, the metrics, the experimental settings, and the obtained results.

4.1 Document Collections

We used two different sets of documents containing both local and global examples. These document collections were:

- **Topix news articles collection** : We crawled a set of 100 news articles from the Topix website[6], which includes many regional news. Each document was then manually classified using the location relevance criteria that was previously introduced, resulting in a collection containing 50 local documents and 50 global documents.
- **ODP Web pages collection** : We experimented with a collection of 8000 Web pages containing at least one geographic reference, 4000 classified as *local* and 4000 classified as *global*, randomly crawled from the Open Directory Project[7] (ODP). Pages under small locations in the "Regional" portion of the directory were regarded as local (i.e., US cities and US states), while pages outside the "Regional" category or under a large region (i.e., USA) were regarded as global.

4.2 Evaluation Metrics

We considered standard information retrieval evaluation metrics to assess the performance of the various classifier configurations. Thus, precision and recall, were computed for both the local and global document classes. Precision (P) is the ratio of the number of items correctly assigned to the class divided by the total number of items assigned to the class. Recall (R) is the ratio of the number of items correctly assigned to a class as compared with the total number of items in the class. Since precision can be increased at the expense of recall, we also compute the F1 measure, which combines precision and recall into a single number using the formula $F1 = 2PR/(P + R)$.

Given the binary nature of our classification problem, we also measured results in terms of accuracy and error. Accuracy is the proportion of correct results (both true positives and true negatives) given by the classifier. Error, on the other hand, measures the proportion of instances incorrectly classified, considering false positives plus false negatives.

4.3 Results and Discussion

For both the news and the ODP document collections, we trained SVM classifiers using different combinations of the proposed features. For each case, a 10-fold cross validation was performed. The considered feature combination experiments are presented in Table 1. Tables 2 and 3 overview the results.

By looking at the results, we can see that, in both collections, the textual features, as well as the locative features, are by themselves able to provide relatively high accuracies. Nonetheless, by combining them we can improve the accuracy by more than 8% on the ODP collection and 3% on the Topix collection. High level locative features had worse results than we had anticipated. We think this might be related with a relatively poor effectiveness of both Web-a-Where and GIPSY in detecting the correct geographic scope. In a seperate study we showed that

[6] `http://www.topix.com`
[7] `http://www.dmoz.org`

TW	Word stems derived from the entire document.
TK	Word stems generated from the key words.
TWK	Combination of the features from cases TW and TK.
U	URL n-grams.
LS	Simple locative features.
LH	High level locative features.
LSH	All the locative features.
T+L	The best textual feature, plus the best locative feature.
T+L+U	The best textual and locative features, plus the URL n-grams.

Table 1. The feature combinations considered in our experimental setup.

	Precision		Recall		F-Measure			
Experiment	Local	Global	Local	Global	Local	Global	Error	Accuracy
TW	0.81	0.79	0.79	0.81	0.8	0.8	19.86	80.14
TK	0.74	0.74	0.74	0.74	0.74	0.74	25.86	74.14
TWK	0.82	0.8	0.8	0.82	0.81	0.81	19.09	80.92
U	0.66	0.69	0.72	0.63	0.69	0.66	32.47	67.53
LS	0.78	0.86	0.88	0.74	0.82	0.8	18.85	81.15
LH	0.56	0.69	0.87	0.3	0.68	0.42	41.31	58.69
LSH	0.58	0.82	0.93	0.31	0.71	0.45	37.58	62.42
T+L	0.89	0.91	0.91	0.88	0.9	0.89	10.43	89.57
T+L+U	0.89	0.91	0.91	0.89	0.9	0.9	10.15	**89.85**

Table 2. Results obtained for the classification algorithm, over the ODP collection.

	Precision		Recall		F-Measure			
Experiment	Local	Global	Local	Global	Local	Global	Error	Accuracy
TW	0.53	1	1	0.1	0.69	0.18	45	55
TK	0.71	0.61	0.48	0.8	0.57	0.69	36	64
TWK	0.54	1	1	0.16	0.7	0.28	42	58
U	0.45	0.39	0.6	0.26	0.51	0.31	57	43
LS	0.63	0.6	0.54	0.68	0.58	0.64	39	61
LH	0.54	0.51	0.24	0.8	0.33	0.62	48	52
LSH	0.67	0.62	0.56	0.72	0.61	0.67	36	64
T+L	0.71	0.64	0.58	0.76	0.64	0.7	33	**67**
T+L+U	0.54	0.65	0.86	0.26	0.66	0.37	44	56

Table 3. Results obtained for the classification algorithm, over the Topix collection.

Web-a-Where and GIPSY assign documents to the correct scopes approximately 50% and 21% of the times, respectively [28].

Using the URL n-grams slightly improves the ODP results. The same did not happen with Topix, but it is understandable, since an information gain analysis of the features showed us that the most significative n-grams were the top level domains like *.gov* and *.us*, which are hardly found in the Topix news.

An information gain analysis of the features used in the classifiers also showed that the simple locative features are the most important, specially the ones that consider the aggregated count of sub-locations. Regarding the textual features, the same analysis pointed words like *hotel*, *restaurant*, *park*, or *hike*, as highly discriminative.

Although not exploited in this work, we believe that the proposed document classification scheme can be used to improve the quality of the results in tasks such as document retrieval or online contextual advertisement. For document retrieval applications, the locality classification can be pre-computed off-line, since it is query-independent. At query time, given that we can also classify queries as either local or global (for instance using the technique proposed by Gravano et al. [3]), we can re-rank the results so that more global or local documents are ranked higher in the results list shown to the user. The paper by Gravano et al. already provided a thorough discussion on similar ideas. For contextual advertisement, at run-time, we can attempt to localize the advertisements to either the geographical area discussed in the document (particularly interesting to pages where we have a high confidence in that they are local), or to the area of the user that is accessing the document, estimated for instance through the user's IP address (more interesting for less local pages, or for global pages).

In the context of Geographic Information Retrieval, Cai extended the vector space model in order to seperately consider a thematic similarity and a geographic similarity [27]. The confidence score produced by our SVM classifier might be a valid option for weighting these similarities in an overall formula.

5 Conclusions and Future Work

This paper presented a locational relevance categorization scheme for textual documents. We discussed how documents can be represented through features based on textual terms, URL n-grams and place references, extracted through text mining, for the purposes of determining their locational relevance. Using these features, a Support Vector Machine classifier for automatically determining the locational relevance was tested. Empirical results indicate that, for many documents, locational relevance can be determined effectively. We also compared different combinations of features to determine their impact on classification effectiveness. Results showed that using feature vectors that combine weighted text terms with features related to the locality of the document (e.g. place names extracted through text mining) results in increased performance.

Several challenges remain for future work. Assuming three document classes (i.e., global *G*, local *L*, and somewhat local *SL*), instead of our binary classification, we could design up to four separate classification tasks, depending on application needs: (1) to discriminate between classes L,SL and G; (2) to discriminate between classes L and SL,G; (3) to discriminate between classes L and G, ignoring the SL classes; (4) to simultaneously discriminate between all the three classes. This may help in dealing with the difficult classification cases.

We also plan to experiment with additional features in the classifier. In the context of search engine queries, Jones et al. studied the relationship between the non-location part of an explicit geographic query and the distance of the query's location part from the issuer's IP location [25]. They found that geographic queries have a varied distance distribution and, therefore, different localization capabilities. In the context of Web documents, we can also use the distance or

the area of overlap between the area corresponding to the Internet address of the server hosting the document and the geographic scope of the document, as an additional feature for classifying documents as either *local* or *global*.

Moreover, some of the characteristics that make a document either local or global may not be directly observable in the document itself, but rather in other contextual information related to the document. Previous research on Web document classification has shown that better performance can be achieved trough combinations of content-based features with additional features derived from the neighboring documents in the link structure of the web graph [21,22]. Previous experiments dealing with geographic context information have already accounted with similar ideas, as for instance Gravano et al. [3], in classifying search engine queries as either local or global, used a sample of the search results returned for a given query rather than the words of the query itself. For assigning geographic scopes to Web documents, Ding et. al proposed to use the distributional characteristics of the locations associated with HTML in-links [1]. It would be interesting to integrate, into our feature vectors, information about the distributional characteristics of locations in related documents, having this notion of relatedness coming from either textual similarity or from linkage information. Our currently ongoing work is addressing these ideas, aiming at the application of locational relevance classifiers in geographical IR and contextual advertising.

References

1. 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, 545-556.
2. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004) Web-a-where: geotagging web content. In Proceedings of the 27th international ACM SIGIR Conference on Research and Development in information Retrieval, 273-280.
3. Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. (2003) Categorizing web queries according to geographical locality. In Proceedings of the 12th international Conference on information and Knowledge Management, 325-333.
4. Zhuang, Z., Brunk, C., and Giles, C. L. (2008) Modeling and visualizing geo-sensitive queries based on user clicks. In Proceedings of the 1st international Workshop on Location and the Web, 73-76.
5. Witten, I. H., and Frank, E. (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco
6. Woodruff, A. G. and Plaunt, C. (1994) GIPSY: Automated geographic indexing of text documents. Journal of the American Society for Information Science 45, 9, 645-655.
7. Cristianini, N. and Shawe-Taylor, J. (2000) An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.
8. Johansson, M. and Harrie, L. (2002) Using Java Topology Suite for real-time data generalisation and integration. Proceedings of the 2002 workshop of the International Society for Photogrammetry and Remote Sensing.
9. Leidner, J. L. (2008). Toponym Resolution: a Comparison and Taxonomy of Heuristics and Methods.

10. Yang, Y. (1999) An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1, 1-2, 69-90.
11. Sebastiani, F. (2002) Machine learning in automated text categorization. ACM Computer Surveys 34, 1, 1-47.
12. Joachims, T. (1998) Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning, 137-142.
13. Forman, G. (2003) An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research. 3, 1289-1305.
14. Apté, C., Damerau, F., and Weiss, S. M. (1994) Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12 (3), 233-251.
15. Genkin, A., Lewis, D. D. and Madigan, D. (2004) Large-Scale Bayesian Logistic Regression for Text Categorization. Rutgers University Technical Report.
16. Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., and Mahoney, M. W. (2007) Feature selection methods for text classification. In Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 230-239.
17. Sang, E. T. K. and De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. Proceedings of the 7th Conference on Natural Language Learning, 142-147.
18. Kornai A. (2003) Proceedings of the HLT-NAACL 2003 workshop on the analysis of geographic references.
19. Garbin, E. and Mani, I. (2005) Disambiguating toponyms in news. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 363-370.
20. Rauch, E., Bukatin, M., and Baker, K. (2003) A confidence-based framework for disambiguating geographic terms. In Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, 50-54.
21. Chakrabarti, S., Dom, B., and Indyk, P. (1998) Enhanced hypertext categorization using hyperlinks. In Proceedings of the 1998 ACM SIGMOD international Conference on Management of Data, 307-318.
22. Qi, X. and Davison, B. D. (2006) Knowing a web page by the company it keeps. In Proceedings of the 15th ACM international Conference on information and Knowledge Management, 228-237.
23. Baykan, E., Henzinger, M., Marian, L. and Weber, I. (2009) Purely URL-based Topic Classification. In Proceedings of the 18th international World Wide Web Conference, Alternate Track Papers and Posters, 1109-1109
24. Baykan, E., Henzinger, M., and Weber, I. (2008) Web page language identification based on URLs. In Proceedings of the VLDB Endowment, 1 (1), 176-187.
25. Jones, R., Zhang, W. V., Rey, B., Jhala, P., and Stipp, E. (2009) Geographic intention and modification in web search. International Journal of Geographical Information Science, 22 (3), 229-246
26. Yu, B. and Cai, G. (2007) A query-aware document ranking method for geographic information retrieval. In Proceedings of the 4th ACM workshop on Geographical information retrieval, 49-54
27. Cai G. (2002) GeoVSM: An Integrated Retrieval Model for Geographic Information, GIScience, 65-79
28. Anastácio, I., Martins, B., and Calado, P. (2009) A Comparison of Different Approaches for Assigning Geographic Scopes to Documents. In Proceedings of the 1st INForum - Simpósio de Informática.