A Graph-Ranking Algorithm for Geo-Referencing Documents

Bruno Martins and Mário J. Silva
Departamento de Informática - Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal

E-mail: bmartins@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt

Abstract

This paper presents an application of PageRank for assigning documents with a corresponding geographical scope. We describe the technique in detail, together with its theoretical foundations. Experimental results are promising, comparing favorably with previous proposals.

1 Introduction

Text documents often encode a geographic context. Geo-referencing this data, that is assigning geographic scopes to documents with basis on the geographical references in the text, is a task getting increasing attention [1, 6]. This can be seen as document classification, with classes (geo-scopes) assigned according to the document's degree of locality. However, geo-referencing texts poses many challenges to traditional techniques. For instance, the amount of training data per parameter (i.e. the number of references to a particular geographical concept) is so low that there are no sufficient repeatable phenomena to base probabilistic methods on. Location names are not enough for classification, as even important ones are usually not repeated. With few exceptions, most work in automated classification has also ignored dependencies between classes and/or features. Typical methods treat the items to classify as a "bag-of-features," not accounting for the possible relations that may exist among them (e.g. region containment).

This paper proposes a novel geo-referencing approach, using geographical references from the text and an ontology-based technique for combining them and infer a global scope for each document. This combination relies on PageRank [12], applied to a graph generated from the ontology. Intuitively, our scheme works well because it goes beyond the local context of a document and simple connectivity heuristics, recursively taking into account information drawn from the entire geographic ontology.

Research partially supported by FCT - Fundação para a Ciência e Tecnologia, under grants POSI/SRI/47071/2002 (project GREASE) and SFRH/BD/10757/2002 (PhD scholarship).Special thanks to Daniel Gomes and Marcirio Chaves for their comments on early drafts of this paper.

2 The Geographic Scope of a Document

The recognition of geographical references in text is discussed in a separate publication [10]. Here we consider the set of geographical references already extracted from the document's text and weighted according to occurrence frequency. A core component of our approach is a geographic ontology, which provides both the vocabulary (used in extracting the geographical references) and the relationships among geographical concepts (used to combine and disambiguate the available information in order to assign a scope).

Ambiguity in geographic references is an important caveat, as place names lack precision in their meaning. The same name can be used for several locations (referent ambiguity), and the same location can also have multiple names (reference ambiguity). The ontology and the extraction software have mechanisms for dealing with this (e.g. alternative names and disambiguation heuristics). Most references in the text end up associated with an ontology concept, although some can also remain associated with several concepts. Assigning a scope to each document is the final disambiguation step, combining all the extracted (possibly noisy) information. Although we take a simple approach of associating each document with one encompassing scope (instead of allowing them to be associated with several regions), different degrees of locality are nonetheless considered, whether documents are relevant to a geographically broad audience or to relatively narrow areas.

An initial evaluation on the extraction of geographical references from text showed a good precision score (80.34%), although recall still needs improvement (tests suggest that only as much as 68% of the geographical references are being found). The performance of this extraction stage dictates the performance on assigning scopes, which gives a good motivation for a separate study of both tasks.

2.1 Ontologies of Geographical Concepts

We developed two ontologies for our experiments, merging information from public sources. One concerns global information in multiple languages, while the other focuses on the Portuguese territory [4]. In the Portuguese ontology,

Ontology Statistic	Portuguese	Multilingual
Concepts	418065	12293
Relationships	419072	12258
Part-of relations	418340 (99.83%)	12245 (99.89%)
Equivalence relations	395 (0.09%)	1814 (14.80%)
Adjacency relations	1132 (0.27%)	13 (0.10%)
Avg. ancestors	1.0016	1.0703
Avg. descendants	10.5562	475.4454
Avg. equivalents	1.99	3.8270
Avg. adjacents	3.54	6.5
Without ancestors	3 (0.00%)	1 (0.00%)
Without descendants	374349 (89.54%)	12045 (97.98%)
Without equivalents	417867 (99.95%)	11819 (96.14%)
Without adjacents	417739 (99.92%)	12291 (99.99%)

Table 1. The Geographical Ontologies.

we aimed at covering most geographical names related to this national territory. Experiments in this case deal with referent class ambiguity problems as, for instance, street names are commonly used in other contexts (i.e. person names). On the other hand, experiments with the global ontology deal more with referent ambiguity problems, as different places around the globe share the same name.

A characterization of the data is of special importance. Table 1 shows some statistics. If the graph of ontological relations shares properties with the Web graph, then the PageRank algorithm should provide good results. Stochastic models for the Web have been previously described, and there is also substantial work on the analysis of these models [3]. Our statistics indicate common properties, and the application of the PageRank algorithm to a graph derived from our ontologies should not impose significant problems. First, they are much smaller than typical Web graphs. Computation cost is therefore manageable, since PageRank is quite fast. Various methods have also been proposed for accelerating the algorithm [8, 9], and with similar approaches our technique could scale to even larger ontologies. A large percentage of the relations represent a hierarchical concept organization, which is also reported to occur on the Web (although in a lesser scale [7]). Almost all ontology concepts the are somehow interconnected. Still, a small percentage corresponds to the roots and leafs in the hierarchy, and these concepts are less "related" to others. However, some techniques have been proposed for dealing with these sink effects (also referred to as dangling nodes in the Web), such as adding artificial links to all other nodes [8].

Perhaps the most important characteristic of the ontologies is the heterogeneity of relationships, i.e. places are related to each other in different ways. Traditional methods assume that all the links have the same "endorsement" semantics and are equally important. Directly applying these techniques could result in unreasonable rankings, and we should use specializations that account for different types of relationships between nodes. Because the original PageRank gives the same weight to all edges, nodes with more in-links tend to get higher ranks, whether or not they are the most important to the problem at hand. The use of weighted edges is a good means of mitigating these issues.

3 Graph Ranking Algorithms

Although originally designed to determine the importance of Web pages [12], PageRank has been used in many application domains [2, 5, 11]. There is also a considerable amount of work focusing on all aspects of PageRank, namely stability, convergence speed and memory consumption [9]. Formulating our task as a graph-ranking problem has therefore the advantage of building on a large background of theoretical and experimental research.

3.1 Computing PageRank

Traditionally, PageRank has been computed as the principal eigenvector of a Markov chain probability transition matrix. In this paper, we consider the PageRank linear system formulation and its iterative solution methods. Formally, let $G = (V,E)$ be a directed graph with the set of nodes V and the set of edges E, where E is a subset of $V * V$. For a given node V_i, let $In(V_i) \subset V$ be the set of nodes that point to it, and let $Out(V_i) \subset V$ be the set of nodes that V_i points to. The PageRank of a node V_i is defined as:

$$PR(V_i) = (1-d) + d * \sum_{j \varepsilon In(V_i)} \frac{1}{|Out(V_j)|} PR(V_j)$$

where d is a damping factor set between 0 and 1, representing the probability of jumping from a given node to another random node in the graph. Because the equation is recursive, it must be iteratively evaluated until $PR(V_i)$ converges, that is, the error rate for any node falls bellow a given threshold. The error rate of a node V_i is defined as the difference between the "real" score of the node ($PR(V_i)$) and the score computed at iteration k, ($PR^k(V_i)$). Since the real score is not known a-priori, this error rate is approximated with the difference between the scores computed at two successive iterations, given by $PR^{k+1}(V_i) - PR^k(V_i)$. In practice, convergence is quite fast, particularly in the case of small graphs [11]. In our experiments, we limited the number of iterations to 10, also halting the algorithm if the error rate at a given iteration is bellow 0.001. This gives results of sufficient quality, as we are more interested in ranking order rather than in the actual values. The running time of the algorithm is $O(|V| * I)$, where $|V|$ is the number of nodes and I is the number of iterations until convergence.

The original PageRank definition assumes unweighted graphs but, in many applications, it may be useful to integrate into the model connection strengths between the nodes. We can easily consider weights w_{ij} for the edges that connect nodes V_i and V_j. The original PageRank formulation does not include node weights n_j either, but it was suggested that by changing the random jump parameter to be nonuniform, the resulting algorithm can be bi-

ased to prefer certain nodes [12]. Another approach for modeling node weights consists of adding artificial self-links [2]. More recently, PageRank personalization methods have been proposed, including "source strengths" [5] and query-dependent techniques [13]. These are based on restricting the choice of random jumps, so that certain nodes are preferred to arbitrarily chosen ones. Bellow we show the graph-ranking formula that accounts for edge and node weights when computing the score associated with a node.

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

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

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

It is important to note that the final values obtained with PageRank are not affected by the choice of initialization values, and only the number of iterations required to achieve convergence may be different. When formulating a task as a graph ranking problem, it is therefore essential that all relevant information is expressed in the graph itself, as opposed to assigning meaning to the initialization values.

3.2 PageRank in Geo-Referencing Documents

We propose the use of PageRank (based on formula $S(V_i)$ above) for geo-referencing documents, after geographical references have been extracted from the texts and associated with the corresponding concepts at an ontology. We assume each concept in the ontology corresponds to a possible class (i.e. a geographic scope), although we can think of a formulation where each class maps to a set of concepts and not all concepts are possible classes.

Formally, each document can be represented as a set of features F. Each feature $F_i = (w, N') \in F$ corresponds to a geographical reference from the text, and associates a weight w to a set of concepts N' from the ontology, according to occurrence frequency. The complete ontology is a tuple $O = (C, T, R)$, consisting of a set of concepts C, a set of relation types T and a set of relation statements R. Each $R_i \in R$ assumes the form $R_i = (C_a, T_j, C_b)$ and states that concept $C_a \in C$ has a relationship of type $T_j \in T$ with concept $C_b \in C$. We are interested in assigning each document with a class $C_i \in C$ (i.e. a geographical scope corresponding to a concept at the ontology).

In order to apply PageRank, the set of features F and the ontology O both have to be represented in a graph $G = (V, E)$. First, equivalent concepts at the ontology are collapsed into a single one. Each concept $C_a \in C$ is then represented as a node V_a in the graph G, and each relation statement $R_i = (C_a, T_j, C_b)$ is represented by two directed edges $E_{a,b}$ and $E_{b,a}$. Different types of relations T_i in the ontology correspond to different edge weights in the graph (i.e. equivalence relationships result in collapsed nodes, sub-region-of relationships are given a high weight in the graph, and adjacency relationships are weighted as not much important). The feature weights are used to weight nodes in G, both trough the use of artificial self-links $E_{i,i}$ and a source strength parameter s_i. For "unrelated" concepts C_i at the ontology (i.e. with no sub-region-of ancestors or descendants), and in order to avoids sink effects [2], we also generate artificial edges $E_{i,j}$ with a very small weight, linking to all other nodes $V_j <> V_i$ in the graph.

In the future, we plan on using a systematic approach for tuning the weights, but for now this issue has been left out of our experiments. Other important parameters are the damping factor (in our case set to 0.9), the number of iterations until convergence (limited at 10), and the error rate threshold (set to 0.001). Finally, instead of initializing the algorithm with random values, we give each node the sum of the corresponding feature weights. This accelerates convergence, as many important nodes are already ranked higher.

After ranking scores are computed, we still need to select a geographical scope. This is done by normalizing the results and selecting as scope the "most general" V_i node that is ranked higher above a threshold, or `NULL` if this threshold is not reached. By most general, we mean the node corresponding the common broadest concept at the ontology, obtained from the set of all nodes sharing the highest score.

4 Experimental Results

We started with simple tests over artificially generated data, since this allows evaluating our approach independently of recognizing geographic references in text. The tests corresponded to typical situations were combining the available data could provide the means to disambiguate the geographical scope of a document (e.g. multiple references in a document to geographic concepts that are somehow related, or the presence of "noise" references in a document). In all cases, the algorithm converged in a few seconds to the correct result, often in less than 10 iterations.

To evaluate scope assignments in a realistic scenario, we first needed to obtain pre-classified documents with coherent human-assigned scopes. We used two sources, namely the Reuters-21578 newswire collection, and randomly selected Web pages from the Open Directory Project (ODP), located under `Regional` (about 1,000,000 pages) and the `Regional:Europe:Portugal` (about 100 pages).

In the Reuters collection, and using the global ontology, we achieved an accuracy of 92% when matching the assigned scopes with the country that was given for each document. In many cases we were able to assign scopes with a finner level of detail than countries, but we had no automatic way of evaluating the correction of these assignments. Each document had on average 6 geographical concepts recognized in the text. Using an alternative method of assigning scopes based on the most frequently occurring geographic reference resulted on an accuracy of 76%.

Multilingual global ontology : ODP:Regional		
	Measured Accuracy	
Granularity Level	Most Frequent	Graph-Ranking
Continent	91%	92%
Country	76%	85%
Exact Matches	67%	72%
Portuguese ontology : ODP:Regional:Europe:Portugal		
	Measured Accuracy	
Granularity Level	Most Frequent	Graph-Ranking
NUT 1 (3 regions)	84%	86%
NUT 2 (7 regions)	58%	65%
NUT 3 (30 regions)	44%	59%
Municipalities	28%	31%
Exact Matches	34%	53%

Table 2. Results on Web-pages from ODP.

On the ODP collection, we evaluated scope assignments at different levels of "granularity." Instead of just counting exact matches, we used the scopes as a hierarchical naming scheme, measuring at different levels the number of matches between the our results and the gold standard. The intuition behind is that assigning documents to a corresponding broader region is easier than assigning them to a narrower one. In the case of ODP:Regional, we again used the global ontology and each document had on average 7 geographical references recognized in the text. For ODP:Regional:Europe:Portugal, we used the ontology covering the Portuguese territory and each document had on average 10 geographical references (although these were much more ambiguous). Table 2 shows the obtained results. We can see although precision decreases at higher levels of granularity, results are still of acceptable quality. In the table, NUT stands for "Nomenclature of Territorial Units for Statistics," an European standard for referencing administrative divisions of countries. The graph-ranking algorithm also consistently outperforms assigning scopes with basis on the most frequently occurring reference.

The Web-a-Where system also aimed at finding the "geographical focus" of a Web page, although using different techniques for combining the extracted information from the documents [1]. In pages from ODP, it guessed the correct continent, country, city, and exact scope respectively 96%, 93%, 32%, and 38% of the times. Although we have a good indication that both approaches have similar accuracy, different resources were used in the experiments (e.g. the ontology and the extraction technique), and results can not be directly compared. Other approaches have also been described [6], but comparisons with our work are even harder.

5 Conclusions and Future Work

We proposed a novel approach for automatically identifying the geographic scope of a document, using references extracted from the text, information from an ontology, and a graph-ranking algorithm to combine the available data. Experimental results are promising, showing that our approach compares favorably with previous proposals. Since this method does not require annotated corpora, it is also highly portable across languages and domains, although we still require appropriate geographic ontologies.

Formalizing the problem of geo-referencing documents as a graph-ranking task facilitates the integration of different techniques and information sources, leaving many ideas for future work. This includes tests with other graph-ranking algorithms, using other types of relations between geographic concepts (e.g. adding edges to the weighted graph according to the spatial distance between concepts), and experimenting with different notions of "document," by assigning scopes to either paragraphs or whole Web sites.

References

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web content. In *Proceedings of SIGIR-04, the 27th conference on research and development in information retrieval*, pages 273–280. ACM Press, 2004.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *Proceedings of VLDB-04, the 30th Conference on Very Large Data Bases*, 2004.

[3] G. Caldarelli, P. D. L. Rios, L. Laura, S. Leonardi, and S. Millozzi. A study of stochastic models for the Web graph. Technical Report 04-03, Dipartimento di Informatica e Sistemistica - Universita' di Roma "La Sapienza", 2003.

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

[5] M. J. Conyon and M. R. Muldoon. Ranking the importance of boards of directors. *Management Science*, 2004. (to appear).

[6] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of Web resources. In *Proceedings of VLDB-00, the 26th conference on Very Large Data Bases*, pages 545–556. Morgan Kaufmann Publishers Inc., 2000.

[7] N. Eiron and K. S. McCurley. Locality, hierarchy, and bidirectionality in the Web. In *Proceedings of WAW-03, the 2nd Workshop on Algorithms/Models for the Web Graph*, 2003.

[8] T. Haveliwala. Efficient computation of PageRank. Technical Report 1999-31, Stanford University, 1999.

[9] A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1(3), 2004.

[10] B. Martins and M. Silva. Geographical named entity recognition and disambiguation in Web pages, 2005. (To appear).

[11] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04, the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library, 1999.

[13] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.