

# A Comparison of Different Approaches for Assigning Geographic Scopes to Documents

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.** In this paper, we compare different methods for the automatic assignment of geographic scopes to Web pages, based on place-names mentioned in the text. The methods under study are the Yahoo! Placemaker Web service, the hierarchy-based method originally proposed for the Web-a-Where system, the spatial overlap-based method originally proposed in the GIPSY project, the graph-based method originally proposed in the GREASE project, and three simple baseline methods corresponding to using the most frequently occurring place, the spatial area that covers all mentioned places, or the spatial area that covers all mentioned non-outlier places. The task under study may be included into the broader problem of Geographic Information Retrieval. Experiments were carried out on Web pages from the Regional Section of the Open Directory Project, comparing the automatically assigned scopes against those assigned by human editors. The results show that the Web-a-Where method gives the best results, closely followed by the GraphRank method and by the baseline based on to the most frequent occurring place.

**Sumário** Este artigo compara diferentes métodos para atribuir automaticamente âmbitos geográficos a páginas Web, com base nos nomes de locais mencionados no texto. Os métodos avaliados são o do serviço Web Yahoo! Placemaker, o método proposto para o sistema Web-a-Where baseado em hierarquias, o método proposto no projecto GIPSY baseado na sobreposição espacial, o método proposto no projecto GREASE com base em grafos, e três *baselines* correspondentes a atribuir o lugar mais frequentemente mencionado, a região que cobre todos os locais mencionados, e a região que cobre todos os locais mencionados que não são *outliers*. A tarefa de atribuir âmbitos geográficos pode ser incluída no problema mais abrangente da Recuperação de Informação Geográfica. As experiências utilizaram as páginas Web presentes na secção Regional do Open Directory Project, comparando-se os âmbitos atribuídos automaticamente com os atribuídos por humanos. Os resultados mostram que o método Web-a-Where produz os melhores resultados, seguido de perto pelo GraphRank e pela *baseline* com base no local mais frequente.

**Key words:** Cross-Method Comparison, Geographic Text Mining

---

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

## 1 Introduction

Recently, Geographical Information Retrieval (GIR) has captured the attention of many researchers that work in fields related to text mining and data retrieval. Most text documents can be said to be related to some form of geographic context [9]. However, exploring this information presents non-trivial problems, due to the inherent ambiguity of natural language (e.g., placenames often have other non geographic meanings, different places are often referred to by the same name, and the same places are often referred to by different names). Handling place references in text has nonetheless been addressed by several previous works, with the aim of supporting subsequent GIR processing tasks.

The automatic assignment of geographic scopes to Web documents, based on the place references that are present in the text, is an example of a complex GIR problem that has been getting increasing attention. Given a set of diverse geographic regions, corresponding to the placenames mentioned in a given Web page, the problem concerns finding the geographic region that best summarizes and describes them all. The use of the word *diverse* proposes that the set of regions referenced in a document may not be trivial to combine, since there may be some errors in the disambiguation of the placenames mentioned in the text, and not all of the regions resulting from the disambiguation may correspond to sub-areas of a single well-bounded place, like a city. However, it is assumed that the set is somewhat coherent—we do not expect that some place references correspond to regions in Lisbon, Portugal while other place references in the same document correspond to regions in Sidney, Australia. While several different strategies have been proposed in the past, there is nowadays no clear information about the trade-offs involved in choosing a particular algorithm. Each different algorithm makes specific assumptions, therefore resulting in different approximations for the geographic scope of the documents.

This paper presents an empirical comparison of different methods for the automatic assignment of geographic scopes to Web pages, based on placenames mentioned in the text. The GeoCLEF campaign has addressed the black-box evaluation of GIR systems, some of them considering components for scope assignment [4]. However, to the best of our knowledge, no cross-method comparison on the specific problem of scope assignment has ever been reported. The methods considered in this study are the Yahoo! Placemaker Web service, the hierarchy-based method originally proposed for the Web-a-Where system, the spatial overlap-based method originally proposed in the GIPSY project, the graph ranking method originally proposed in the GREASE project, and three simple baseline methods corresponding to using the most frequently occurring place, the spatial area that covers all mentioned places, or the spatial area that covers all mentioned non-outlier places. The cross-method comparison was carried out by comparing the automatically assigned scopes against those assigned by humans. As reference dataset, we used a collection of 6,000 Web pages extracted from the Regional Section of the Open Directory Project<sup>1</sup>, which contains

---

<sup>1</sup> <http://dmoz.org/>

documents classified according to both broad (e.g., United States) and narrow administrative regions (e.g., Plymouth, a city in Minnesota).

The rest of this paper is organized as follows: Section 2 discusses the recognition and disambiguation of place references in text, a necessary pre-processing step in the task of scope assignment. Section 3 details the different scope assignment approaches that are the target of this study. Section 4 presents the experimental evaluation. Finally, Section 5 presents our conclusions, also giving some ideas for future evaluation studies.

## 2 Recognizing Place References in Text

One of the particular problems that has been extensively explored in the area of GIR relates to handling place references, a.k.a. *geotagging text*. This is a crucial task for other GIR-related problems, such as the determination of the geographic scope of Web pages. Handling place references requires recognizing the mentions to places given over text, by delimiting their occurrences, as well as disambiguating these occurrences into the corresponding locations on the surface of the Earth, by assigning geospatial coordinates to the place references. The main challenges involved in both sub-tasks are related to ambiguity in natural language. Amitay et al. characterized ambiguity problems according to two types, namely *geo/non-geo* and *geo/geo* [3]. Geo/non-geo ambiguity refers to the case of placenames having other non-geographic meanings, since some of very common words are also place names (e.g., Turkey the country, or Reading in England). Geo/geo ambiguity arises when two distinct places have the same name. For instance almost every major city in the Europe has a sister city of the same name in the (so-called) New World.

Leidner surveyed a variety of approaches for handling place references on textual documents [2]. Most methods usually rely on finding the placename in a dictionary of known locations (a *gazetteer*), together with natural language processing heuristics such as default senses (i.e., disambiguation should be made to the most important referent, estimated based on population counts) or geographic heuristics such as spatial minimality (i.e., disambiguation should minimize the bounding polygon that contains all candidate referents). Place reference resolution technology is nowadays mature, and commercial services offering this type of functionalities are starting to appear.

Metacarta<sup>2</sup> is an example of a commercial company that sells state-of-the-art geographic information retrieval technology. The company also provides a freely-available Web service that can be used to recognize and disambiguate place references over text. An early version of the Metacarta geotagger has been described by Rauch et al. [5]. The Yahoo! Placemaker Web service also provides a functionality for geotagging text. The Yahoo! Placemaker is used in all our experiments to handle the recognition and disambiguation of place references, and will be further described in Section 3.1.

<sup>2</sup> <http://metacarta.com/>

### 3 Scope Assignment Approaches

This section presents the different scope assignment methods that are the focus of this study, also discussing particular issues regarding the implementations that were used in the experiments.

#### 3.1 The Yahoo! Placemaker Web Service

Yahoo! Placemaker<sup>3</sup> is a geotagging web service that provides third-party developers the means to enrich their applications or Web sites with geographic information. The service is able to identify, disambiguate, and extract place names from unstructured and semi-structured documents. It is also capable of using the place references in a document, together with a pre-determined set of rules, to discover the geographic scope that best encompasses its contents. Thus, given a textual document, Yahoo! Placemaker returns unique *Where-on-Earth* identifiers (WOEIDs) for each of the named places and scopes. Through these identifiers, one can use the Yahoo! GeoPlanet<sup>4</sup> Web service to access hierarchical information (i.e., containing regions) or spatial information (i.e. centroids and bounding boxes).

There are two flavors of document scopes in Placemaker, namely the geographic scope and the administrative scope. The geographic scope is the place that best describes the document. The administrative scope is also the place that best describes the document, but is of an administrative type (i.e., Continent, Country, State, County, Local Administrative Area, Town, or Suburb). Since the reference document collection that we used for our experiments only contains documents assigned to administrative regions, we limited our cross-method comparison to using Placemaker's administrative scopes.

Placemaker is a commercial product and not many details are available regarding its functioning. However, some information about the service is available in the Web site, together with its documentation. For instance, the Web site claims that when the service encounters a structured address, it will not perform street level geocoding but will instead disambiguate the reference to the smallest bounding named place known, frequently a postal code or neighborhood. The Web site also claims that besides place names, the service also understands geography-rich tags, such as the W3C Basic Geo Vocabulary and HTML microformats<sup>5</sup>. However, no details about the rules that are used in the scope assignment process are given in the documentation for the service.

The Placemaker Web service accepts plain text as input, returning a XML document with the results. The service has an input parameter that allows users to provide the title of the document separately from the rest of the textual contents, weighting the title text as more representative. In our experiments, we used the Web service as a black-box to assign scopes to the Web documents, using the option that weights the title text as more important than the rest.

<sup>3</sup> <http://developer.yahoo.com/geo/placemaker/>

<sup>4</sup> <http://developer.yahoo.com/geo/geoplanet/>

<sup>5</sup> See <http://microformats.org/wiki/geo> or <http://microformats.org/wiki/adr>

### 3.2 The GIPSY Scope Assignment Method

In one of the pioneering works in the area of GIR, Woodruff and Plaunt proposed a technique for computing the geographic scope of a given textual document based on the place references discovered in the text [2]. Their method is based on disambiguating the place references into their respective bounding polygons. The geographic scope of the document is afterwards computed using the overlapping area for all the polygons, trying to find the most specific place that is related to all the place references made in the text.

Consider, for instance, the following example. A given document contains references to Portugal, to the city of Lisbon, and to the Iberian Peninsula. After disambiguation, each of these place references is represented as a bounding box. For the GIPSY algorithm, the bounding boxes are seen as thick polygons, with a base positioned in an  $(x, y)$  plane, but extending upwards a distance of  $z$ , to a higher parallel plane. One by one, the three bounding boxes corresponding to the place references are analyzed by the GIPSY algorithm, in order to build a skyline of bounding boxes. Three different cases can occur:

- When adding the bounding box for Portugal, it would not intersect with other bounding boxes. Portugal would simply be laid at  $z = 0$ ;
- When adding the bounding box for Lisbon, it would be completely contained within a bounding box which already exists on the skyline, in this case the bounding box for Portugal. Lisbon would be laid on top of the Portuguese bounding box, i.e. its base would be positioned at a higher  $z$  plane.
- When adding the bounding box for the Iberian Peninsula, it would intersect other bounding boxes but it would not be wholly contained. The bounding box would first be split into multiple polygons. Then, the intersecting polygons would be laid on top of the existing bounding boxes, one on top of Lisbon and another on top of Portugal, and the non-intersecting polygon (continental Spain) would be laid at a lower level.

Finally, all the bounding boxes would be sorted according to their  $z$  order and the highest ranking bounding box is selected as the scope. In our example, the resulting scope would correspond to the area of Lisbon.

The Yahoo! Placemaker Web service was used to recognize place references in the documents and disambiguating them into bounding boxes. The Java Topology Suite [8] provided the required functionality related to spatial computations.

### 3.3 The Web-a-Where Scope Assignment Method

In the context of the Web-a-Where project, Amitay et al. proposed a technique for assigning Web documents to the corresponding geographic scopes [3]. Their technique leverages on part-of relations among the recognized place references, provided by a hierarchical gazetteer. The basic idea is that, for instance, if several cities from the same country are mentioned, this might mean that this country is the scope, i.e. the algorithm tries to generalize from the disambiguated place references. More specific places are scored higher if they are the only places

mentioned, but at the same time we also permit a general region to be chosen if several different places in it are mentioned, with no specific emphasis on any.

The algorithm starts by placing the recognized place references in a locational hierarchy. By looping over the disambiguated references, the algorithm aggregates the importance of the various levels in the hierarchy. The levels are then sorted by importance and the highest ranked level is returned as the scope.

Consider, for instance, the following example. After place reference disambiguation, a document contains a reference to the city of Lisbon (i.e., **Europe/Portugal/Lisbon**) with confidence 0.5, and a reference to Portugal (i.e., **Europe/Portugal**) with confidence 0.8. A taxonomy is first built from the disambiguated place references. Then, the taxonomy nodes are weighted according to the disambiguated place references, where **Europe/Portugal/Lisbon** gets a weight of  $0.5^2$  and **Europe/Portugal** gets a weight of  $0.8^2$ . This quadratic scoring function increases the relative weight of very confident disambiguations.

After assigning the initial weights we propagate scores to the parent regions in the taxonomy, by adding the scores of their respective sub-regions. Thus, we add  $0.5^2 \times 0.7$  to **Europe/Portugal** and  $0.5^2 \times 0.7^2 + 0.8^2 \times 0.7$  to **Europe**. The multiplying discount parameters correspond to those originally reported in the Web-a-Where paper. Finally, we select the highest scoring taxonomy node as the scope to assign (Portugal, in our example).

We used the Yahoo! Placemaker Web service as the means for recognizing place references in the documents and disambiguating them into nodes in the hierarchical gazetteer used in the Yahoo! GeoPlanet platform.

### 3.4 The GraphRank Scope Assignment Method

In the context of the GREASE project, Martins and Silva proposed a scope assignment method based on a graph-ranking approach [6]. The idea was to represent the gazetteer used for place reference disambiguation as a graph, where the nodes correspond to different places and the edges correspond to semantic relationships (*part-of*, *containment* or *adjacency*) between places. Nodes on this graph can be weighted according to the occurrence frequency of place references in a document, and edges can be weighted according to the relative importance of the different types of relationships. A graph-ranking algorithm, *PageRank*, is then applied to this graph, and finally the highest ranked node is selected as the scope. In case of ties, the node connected to the highest number of edges is selected. By propagating scores across the graph, this algorithm tries, at the same time, to generalize and to specify from the available information, in order to find the region that best reflects the scope of the document. For computing the PageRank score, we used the open-source weighted PageRank implementation made available by the Laboratory for Web Algorithmics of the University of Milan [7]. Since this implementation does not allow for weighted nodes, we instead use self-edges, one for each occurrence of a given place in the document.

Consider the following example. A given document contains references to United States and to Los Angeles, which are extracted and disambiguated with confidences of 0.9 and 0.8, respectively. In order to generate the graph, we would

first find the hierarchical parents of the references that are made in the document, the neighboring places to the document references, and the hierarchical parents for these neighboring places. The places discovered through the above procedure would be the nodes of the graph and the relationships between them would be used to produce directed edges between the nodes. For all the nodes with no outlinks (i.e., the roots and the leaf nodes) we would add artificial edges to all other nodes in the graph. The part-of, containment, and adjacency edges would all get a value of 0.4, and artificial edges 0.01. These weights were tuned empirically, and a challenge for future work consists of using automated approaches for tuning these parameters. United States and Los Angeles would also have edges to themselves, with weights equal to their confidence scores. The PageRank algorithm would then be applied and, in the end, the highest scoring node would be selected as the scope.

The Yahoo! Placemaker Web service was used for recognizing place references in the documents and disambiguating them into nodes in a hierarchical gazetteer. The complementing Yahoo! GeoPlanet Web service was used to retrieve the parent and neighboring regions for each of the place references that are made in the document. The graph is built by considering all this information.

### 3.5 Baseline Scope Assignment Methods

The previously described methods make non-trivial assumptions about how place references should be combined to discover the geographic scope of a document. In order to assess what are the gains introduced by these assumptions, we implemented three simple baseline methods, which we now describe:

- **Assigning the scope according to the most frequently occurring place reference** - The number of times a place is referenced in a document reflects the importance of that place to the document's subject. We therefore experimented with a simple scope assignment method that chooses the most frequently occurring place reference as the scope. In case of ties, the place reference corresponding to the largest area is chosen.
- **Assigning the scope according to the bounding box that covers all the place references** - The different place references made in the document should all contribute to the document's scope. We therefore experimented with a simple scope assignment method that computes the bounding box that covers all the place references made in the document.
- **Assigning the scope according to the bounding box that covers all the place references that are not outliers** - This is a refinement of the previous strategy, in the sense that not all place references should contribute to the scope, but only the place references that are somewhat interrelated. The idea is to be able to filter the errors made while recognizing and disambiguating place references, as well as filtering out the place references that are only tangential to the content of the document. We first compute the average centroid point for all the place references made in the document, as well as the average distance between the place references and this centroid.

	All Pages	Countries	States	Cities
Number of documents	6000	2000	2000	2000
Number of ODP sub-classes	1440	692	665	303
Number of different regions	1127	1	51	1075
Average document length (bytes)	4143	4235	3841	4354
Average number of place references	9.2	9.1	9.1	9.4

**Table 1.** Statistical characterization for the test collection of ODP documents.

Then, we filter out those place references whose centroid is at a distance that is greater than twice the average distance value. Finally, we assign a scope corresponding to the bounding box that covers all the remaining place references, if none the closest is chosen. This baseline is inspired on a technique proposed by Smith and Crane for placename disambiguation [10].

We implemented the above three strategies by using the Yahoo! Placemaker Web service to recognize place references in the documents and disambiguate them into bounding rectangles. The Java Topology Suite [8] provided the required functionality related to spatial computations.

## 4 Comparative Experiments

In this section, we describe the details of our empirical evaluation. This includes the experimental design, the datasets, and the evaluation metrics that were considered, as well as the results of the experiments that evaluate the effectiveness of the methods under study.

### 4.1 Dataset

We evaluated the algorithms described in Section 3 for assigning documents to geographic scopes by comparing their assignments to those of the human editors of the Open Directory Project (ODP). Specifically, we selected a random sample of 6,000 Web-pages from the ODP’s `Regional/North_America/United_States` section, that were written in English, were larger than 2 KBytes, and contained at least one geographic reference. The collection contains documents classified according to both broad (e.g., United States) and narrow (e.g., Plymouth, Minnesota) administrative regions. Table 1 presents a statistical characterization of the test collection, separating the pages according to the type of geographic scope to which they belong (i.e., country, state, and city).

### 4.2 Evaluation Metrics

We propose to compare the different approaches by measuring the distance and the relative overlap between the geographic scope that was assigned by the algorithms and the geographic scope that was assigned by the human editors of

	Avg. Distance	Std.dev. Distance	Avg. Overlap	Std.dev. Overlap	Accuracy (D=0 Km)	Accuracy (D<100 Km)	Accuracy (O>0.75)
Placemaker Admin.	1030	<b>1460</b>	0.42	0.49	0.37	0.45	0.39
Web-a-Where	<b>955</b>	1890	0.48	0.49	<b>0.47</b>	0.54	0.47
GIPSY	1265	2247	0.25	0.41	0.14	0.4	0.19
GraphRank	1083	1955	0.48	0.49	<b>0.47</b>	0.53	<b>0.48</b>
Covering Area	2655	3009	0.25	<b>0.38</b>	0	0.21	0.18
Most Frequent	1093	2331	<b>0.49</b>	0.49	0.37	<b>0.55</b>	0.43
Non-outliers	1740	2826	0.36	0.46	0.24	0.39	0.34

**Table 2.** Comparison of human-assigned versus automatically assigned scopes.

the Open Directory Project. Besides the average distances and overlaps, we also compute their standard deviation. Thresholding the distance results at different levels, we also measure results according to the standard information retrieval metric of accuracy (i.e., the proportion of correct results given by the algorithm).

### 4.3 Comparison Results

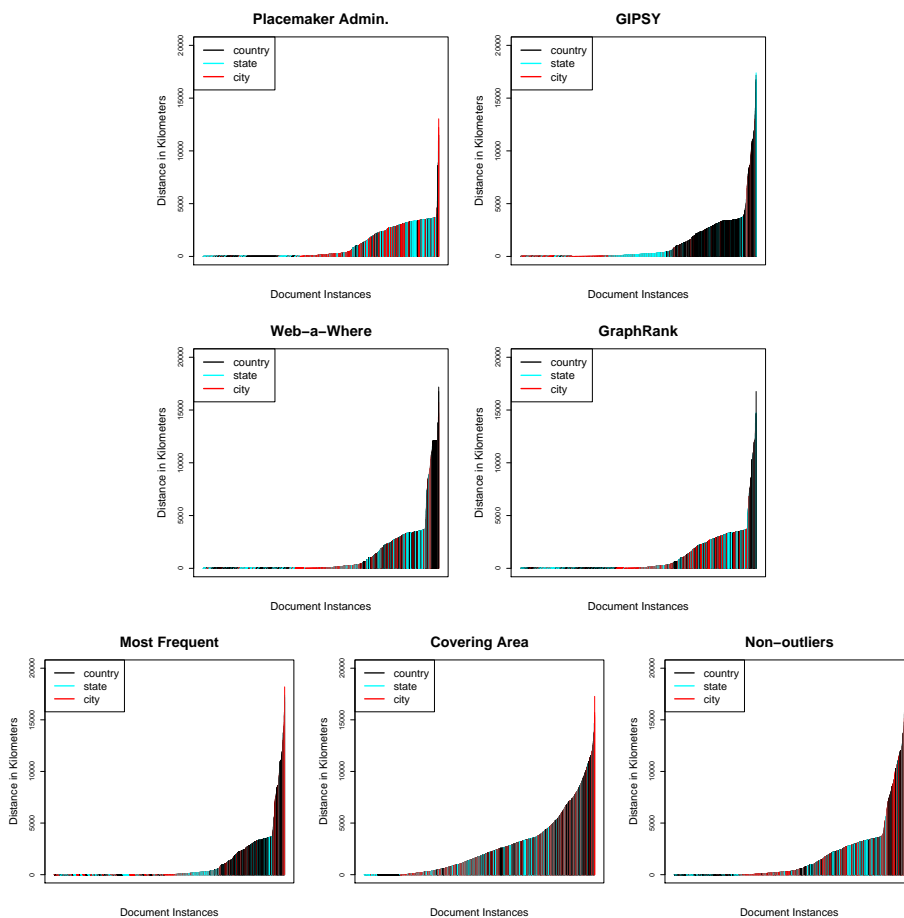
Figure 1 and Table 2 show the obtained results for the cross-method comparison. Figure 1 plots, for each algorithm, the distances from the assigned scopes to the real scopes of the documents. Each value on the  $x$  axis corresponds to a document in the collection, and documents are sorted according to the distance. Table 2 summarizes the values obtained using all the evaluation measures.

The charts show that the covering area baseline produces more errors. In most of the considered approaches, more than half of the documents are assigned to a nearby scope. The GIPSY method and the most frequent baseline produce more errors on pages whose scope corresponds to countries, whereas the other methods have errors equally distributed across countries, states, and cities.

The Web-a-Where method produced the best overall results. The average distance and accuracy to the correct scope were 955 Km and 47%, respectively. The baseline considering the most frequent place reference provided very competitive results, outperforming all other approaches when assigning scopes with an error below 100 Km. The baseline with the most frequent place also obtained the best average overlap with the correct scope (0.49). The GraphRank method did well, matching Web-a-Where’s accuracy for exact matches and obtaining the best accuracy for approximate overlaps.

Since the methods with best performances exhibit different types of problems (e.g., the most frequent baseline fails on countries), their combination seems like a promising approach. For instance, Web-a-Where tends to fail when incorrectly generalizing to a broader area, while the most frequent baseline does not generalize when it should.

We also analyzed how accuracy varies according to the number of place references contained in the documents. Figure 2 presents the obtained results, showing that the errors increase with the number of different place references contained in the document. The baseline method corresponding to the covering area is particularly sensitive to this parameter.



**Fig. 1.** Distances between the human-edited and automatically assigned scopes.

When interpreting the results, it should be noted that the scope assignment algorithms use the information provided by the Yahoo! geotagger about the individual places mentioned in the text. Thus, any errors made by the geotagger influence the outcome of the scope assignment methods (i.e., this test only evaluates geotagging plus scope assignment as a whole). Nevertheless, this does not invalidate the goal of the experiments, which is to compare the scope assignment methods under the same conditions.

## 5 Conclusions and Future Work

In this paper, we compared different methods for the automatic assignment of geographic scopes to Web pages, based on placenames mentioned in the text.

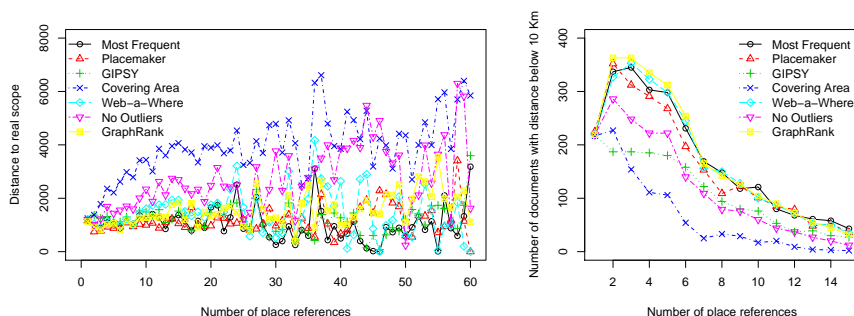


Fig. 2. Correlation between number of place references and scope accuracy.

This is an important pre-processing stage for geographic IR applications. Experimental results showed that overall the Web-a-Where method gives the best values. Nonetheless, this method is closely followed by a simple baseline that assigns the most frequent place reference as the geographic scope, as well as by the graph-based approach.

To the best of our knowledge, this was the first survey and cross-method comparison made in this particular domain. While our results are interesting, there are also many open questions. Below, we list what we consider to be the most interesting paths for future work.

- Optimizing the parameters used in some of the considered approaches. For instance, Web-a-Where uses a hierarchical discount parameter and GraphRank uses different weights for each of the geographic relationship types. Ideally, these parameters should be optimized according to a principled approach (e.g., through simulated annealing, genetic programming, or other optimization methods [11]).
- Devising particular tests to see if some of the algorithms are better for some types of documents. In our experiments, we have seen that all algorithms are similarly affected by the number of place references made in the documents. However, it remains to be seen if the algorithms are all equally robust to the ambiguity problems that may arise in different types of documents. If indeed it is the case that some algorithms are better than others in particular cases, then perhaps we can explore machine learning approaches to select the most appropriate scope assignment method to use in each case. Taking the best algorithm for each document in the test collection, we would get an accuracy of 70%, showing that this is indeed a promising alternative.
- Devising strategies for combining the results produced by the different scope assignment methods into a single scope. Our experiments showed significant differences in the algorithms and it would be interesting to see if a combination of the best algorithms could lead to better results. Possible combination

strategies include taking the bounding box that covers all scopes, or taking the bounding box from their overlapping area.

- Devising experiments to measure the trade-offs in selecting more than one scope for each document. Many documents can not be naturally summarized into a single scope, and some applications may also benefit from having multiple scopes. When considering multiple scopes, there is a chance that we are increasing recall. Precision would nonetheless be impacted, since we would select more scopes incorrectly. A particular challenge is finding the appropriate thresholds for selecting candidates as scopes.
- Devising experiments for measuring how the quality of place reference disambiguation influences the quality of the scope assignments. Instead of always relying on a single approach for place reference disambiguation, it would be interesting to experiment with other geotagging approaches, measuring their performance and their impact on the scope recognition performance. It would also be interesting to see if the scope assignment methods are robust enough to work with all the possible place referents, therefore dispensing the pre-processing step of place reference disambiguation.

## References

1. Woodruff, A. G., and Plaunt, C. (1994) GIPSY: automated geographic indexing of text documents. *Journal American Society Information Sciences*, 45(9).
2. Leidner, J. L. (2007). *Toponym Resolution: a Comparison and Taxonomy of Heuristics and Methods*. PhD Thesis, University of Edinburgh.
3. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004) Web-a-Where: geotagging web content. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*.
4. Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Sanderson, M., Santos, D. and Womser-Hacker, C. (2008) An evaluation resource for geographic information retrieval. In *Proceedings of the 6th Language Resources and Evaluation Conference*.
5. 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*.
6. Martins, B., and Silva, M. J. (2005) A Graph-Ranking Algorithm for Geo-Referencing Documents, In *Proceedings of the 5th IEEE International Conference on Data Mining*.
7. Boldi, P., Santini, M., and Vigna, S. (2005) PageRank as a function of the damping factor. In *Proceedings of the 14th International World Wide Web Conference*.
8. Johansson, M., and Harrie, L. (2002) Using Java Topology Suite for real-time data generalization and integration. In *Proceedings of the 2002 workshop of the International Society for Photogrammetry and Remote Sensing*.
9. Jones, R., Zhang, W. V., Rey, B., Jhala P., and Stipp E. (2008) Geographic intention and modification in Web search. *International Journal of Geographical Information Science*, 22(3).
10. Smith, D. A. and Crane, G. (2001) Disambiguating Geographic Names in a Historical Digital Library. In *Proceedings of the 5th European Conference on Research and Advanced Technology For Digital Libraries*.
11. Spall, J. C. (2003) *Introduction to Stochastic Search and Optimization*, Wiley-Interscience.