Handling Implicit Geographic Evidence for Geographic IR

Nuno Cardoso
University of Lisbon
Campo Grande
1749-016 Lisboa, Portugal
ncardoso@xldb.di.fc.ul.pt

Mário J. Silva
University of Lisbon
Campo Grande
1749-016 Lisboa, Portugal
mjs@di.fc.ul.pt

Diana Santos
Linguateca, SINTEF ICT
Pb 124, N-0314
Oslo, Norway
Diana.Santos@sintef.no

ABSTRACT

Most geographic information retrieval systems depend on the detection and disambiguation of place names in documents, assuming that the documents with a specific geographic scope contain explicit place names in the text that are strongly related to the document scopes. However, some non-geographic names such as companies, monuments or sport events, may also provide indirect relevant evidence that can significantly contribute to the assignment of geographic scopes to documents. In this paper, we analyze the amount of implicit and explicit geographic evidence in newspaper documents, and measure its impact on geographic information retrieval by evaluating the performance of a retrieval system using the GeoCLEF evaluation data.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Design

Keywords

Geographic Information Retrieval, Named Entity Recognition, Wikipedia Mining

1. INTRODUCTION

This paper characterizes the implicit geographic evidence that can be extracted from documents, and evaluates its contribution to the retrieval performance of geographic information retrieval (GIR) systems. We divide the geographic evidence in documents in two types: i) *explicit* geographic evidence, consisting of place names that are normally used to designate geographic areas, such as countries, divisions or territories (for example, "Portugal", "New York City"), and ii) *implicit* geographic evidence, consisting of other named entities that do not refer explicitly to a place name, but are strongly related to a given geographic area, such as monuments, buildings, company headquarters or summits (for example, "Eiffel Tower", "IBM Headquarters" or "CIKM 2008").

Typical GIR systems assume that a document with place names, such as "New York," is also likely to be a document whose geographic areas of interest (or geographic scope) includes the New York City limits. Consequently, it would be also relevant to users searching for a specific topic in this specific geographic scope (for example, querying for "Buildings in New York"). Nonetheless, a document presenting details about the Empire State Building can also be strongly relevant for the same query. Moreover, the lack of explicit references to New York City does not make it irrelevant to the query "Buildings in New York".

Our hypothesis is that, if implicit geographic evidence is present in the documents, and if it is also strongly related to the document scopes, GIR systems can profit from this additional and yet unexplored geographic information. Implicit geographic evidence can therefore be used to reinforce the explicit geographic evidence, or even replace it when there are no place names on the document to derive a document scope within a certain confidence level.

The hypothesis is validated in the paper through several experiments performed on a prototype GIR system enriched with a new Wikipedia-based named entity recognition system, REMBRANDT, capable of capturing and disambiguating multiple types of named entities from text, and extract implicit geographic evidence from the captured named entities. The evaluation is done by measuring the retrieval results on the evaluation data available from past editions of GeoCLEF, an evaluation track for GIR systems organized by CLEF (www.clef-campaign.org).

2. CAPTURING GEOGRAPHIC EVIDENCE

In order to test the hypothesis at stake, we developed REMBRANDT, a named-entity recognition (NER) system capable of classifying named entities (NE) for texts in Portuguese and English, using the 9 main categories and 47 sub-categories defined by the second edition of HAREM, an evaluation contest for Portuguese NER [3]: PERSON, ORGANIZATION, PLACE, DATETIME, VALUE, ABSTRACTION, EVENT, THING and MASTERPIECE. REMBRANDT can handle vagueness in named entities, by tagging the named entities with more than one category or sub-category.

REMBRANDT uses Wikipedia as a raw knowledge source, and explores the Wikipedia document structure to classify all kinds of named entities in the text. By using Wikipedia, REMBRANDT obtains additional knowledge on every named entity that can be useful for understanding the context, detecting relationships with other named entities, and using this information to contextualize and classify surrounding named entities in the text.

A practical application is the use of the Wikipedia page categories to derive implicit geographic evidence for each named entity. REMBRANDT handles category strings as text sentences and searches for place names in a similar way as it is performed on normal texts, generating a list of captured place names that are considered as implicit geographic evidence for the given named entity.

We can therefore divide the NEs found by REMBRANDT on the documents into three levels of eligibility of geographic evidence:

1. **Explicit geographic evidence**, given by NEs that refer to administrative places, such as countries or cities, and landscape places, such as islands or rivers.

2. **Implicit geographic evidence**, given by NEs that refer to organizations, such as city halls, schools or companies, to organized events that take place in a defined place, such as olympic games, rock concerts or conferences, or buildings, such as monuments or fountains.

3. **No geographic evidence**, given by NEs that refer to numbers, dates, persons, abstractions or generic objects.

3. EXPERIMENT AND RESULTS

A prototype GIR system was built to experiment with different weights for implicit and explicit geographic evidence and measure their contribution on the retrieval performance for geographic-flavored search topics. The GIR system and the observed results are described in more detail in a paper presented at GeoCLEF 2008 [2].

REMBRANDT annotated the English and Portuguese ad-hoc collections used in the GeoCLEF tracks since 2005, thus providing the geographic signatures of each document, encompassed as the list of NEs with any geographic evidence. Both collections comprise plain text newspaper articles dated from January 1994 to December 1995. We used the GeoCLEF topics from 2006 until 2008, and we defined different weights to the 3 indexed fields: i) `text`, as in classic IR term index, ii) `explicit local`, containing terms from NEs considered as explicit geographic evidence, and iii) `implicit local`, containing terms from the place names associated to the NEs considered as implicit geographic evidence.

Table 1 presents the best MAP results for the classic IR retrieval (using only the `text` field) and the GIR retrieval (using the three index fields). It shows that the overall GIR retrieval results improve with the use of explicit geographic evidence in some experiments. However, implicit geographic evidence did not always produce the same effect, which means that the proposed process of extraction of implicit geographic evidence needs further optimization before being incorporated in geographic retrieval.

PT	Classic IR				Geographic IR				MAP Diff.
	text	expl. loc.	impl. loc.	MAP	text	expl. loc.	impl. loc.	MAP	
2006	1.0	0.0	0.0	0.1613	2.0	0.25	0.0	**0.1810**	12.2%
2007	1.0	0.0	0.0	0.2730	2.5	0.25	0.0	**0.3037**	11.2%
2008	1.0	0.0	0.0	0.2233	4.0	0.25	0.0	**0.2301**	3.0%
EN									
2006	1.0	0.0	0.0	0.2158	2.25	0.5	0.25	**0.2442**	13.2%
2007	1.0	0.0	0.0	0.2238	2.0	0.5	0.0	**0.2713**	21.2%
2008	1.0	0.0	0.0	0.2528	2.75	0.25	0.0	**0.2630**	4.0%

Table 1: MAP values for IR and GIR retrievals.

4. FINAL REMARKS

In this paper, we wanted to capture and use implicit geographic evidence, along with explicit geographic evidence, to help GIR performance. We used our GIR system to perform such experiments and we analyzed the contribution of each kind of geographic evidence on the retrieval performance of a GIR system.

The amount of explicit and implicit geographic evidence proved to be considerable on the newspaper collections used: an average of 7.7 and 12.7 NEs per document as explicii geographic evidence for Portuguese and English, respectively, and 5.1 and 23.8 for implicit ones. We observed moderate improvements on the GIR retrieval when using explicit geographic evidence, but we did not observe improvements when using implicit geographic evidence.

Although these results do not support our initial belief that GIR retrieval would profit from the inclusion of implicit geographic evidence on the geographic retrieval, a more fine-grained analysis of the results also revealed that REMBRANDT's naïve approach for capturing implicit geographic evidence is not as precise as expected, specially for vague named entities, which generated a considerable amount of unrelated place names and lead to the generation of noisy geographic signatures.

Considering the low precision of REMBRANDT's strategy for extracting implicit geographic evidence, we plan to improve REMBRANDT in order to extract other geographic information present on Wikipedia documents, such as place names in the first paragraph, infobox informations (similar to [1]), or geographic coordinates that normally label many pages about named entities that are not explicitly places. For instance, the Wikipedia page of the Belém Tower (`en.wikipedia.org/wiki/Belém_Tower`) has the coordinates 38° 41′ 29″ N, 9° 12′ 57″ W, and the first paragraph refers that is "located in the Belém district of Lisbon, Portugal". The first paragraph and the geographic coordinates provide more fine-grained information than the Wikipedia category "Buildings and structures in Lisbon," we conjecture that REMBRANDT will therefore be able to generate more precise implicit geographic evidences than the current approach.

5. REFERENCES

[1] S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In Franconi et al., editors, *Proceedings of European Semantic Web Conference (ESWC'07)*, number 4519 in LNCS, pages 503–517. Springer, 2007.

[2] N. Cardoso, P. Sousa, and M. J. Silva. The University of Lisbon at GeoCLEF 2008. In F. Borri et al., editors, *Working notes of CLEF 2008*, Aarhus, Denmark, 17-19 September 2008.

[3] D. Santos, N. Seco, N. Cardoso, and R. Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In N. Calzolari et al., editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, pages 1986–1991, Genoa, Italy, 22-28 May 2006.