The University of Lisbon at GeoCLEF 2008

Nuno Cardoso, Patrícia Sousa and Mário J. Silva
University of Lisbon, Faculty of Sciences, LaSIGE
1749-016 Lisboa, Portugal
{ncardoso,csousa}@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt

Abstract

This paper reports the participation of the XLDB team from the University of Lisbon at the 2008 GeoCLEF task. We focused on developing a better text annotation tool for geo-parsing the documents, handling both explicit geographic evidence (as given by placenames) and implicit geographic evidence (as given by monuments, for example). The query processing and geographic ranking approaches were redesigned to handle thematic and geographic criteria of each search in a non-segregation way. We detail the GIR system, describe the optimisation procedure that preceded the run generation, and dissect the results.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic IR, Named Entity Recognition, BM25 Optimisation, Query Expansion, GeoCLEF, Evaluation

1 Introduction

This paper presents the participation of the XLDB team from the University of Lisbon at the 2008 GeoCLEF task. We matured the ideas implemented on last year's GIR system [3], that achieved interesting results and pointed out some weaknesses of our approach. The main conclusions drawn were the following:

Best practices for handling thematic and geographic criteria. Our GIR methodology so far was moulded on the assumption that the thematic and geographic facets of documents and queries were complementary and non-redundant [1]. We therefore focused our research on GIR prototypes with separated pipelines for handling thematic and geographic subspaces, computing two different ranking scores that were combined in the end to generate a final ranking score. The evaluation results did not show a significative improvement compared to classic IR retrieval, and we wonder whether if this segregational approach is indeed a good practice for GIR [4].

Capturing additional geographic evidence from documents. People describe places of their interest in several other ways, other than explicit placenames. Entities such as "Big Apple", "Kremlin" or "UE Headquarters" are easily connotated to their respective locations, and these entities might have a decisive role on the defining the geographic scope (that is, the geographic area of interest) of the

document. Our shallow text mining approaches often failed to capture essential geographic evidence to geo-reference many documents, and this naïve text mining approach was reflected on poor retrieval results for some geographically challenging topics [3]. We therefore need to reformulate our text annotation tools, making it capable of recognising all kinds of entities with a geographic flavour and grounding them to their corresponding locations.

Smoothing the effects of text and geographic query expansion. Query expansion (QE) is known to improve IR performance in most queries, but often at the cost of degrading the performance on other queries [8]. QuerCol, the QE module used in our GIR prototypes, does not assign weights to the query terms, so the expanded terms have the same weight than the initial query terms. This means that we do not control the impact of QE in some topics, which led in some cases to query drifting and thus lead to poor retrieval results. We want to improve QueCol to perform automatic re-weighting of text and geographic terms, in order to soften the QE effect and prevent query drifting.

For this year's participation, we addressed these topics on the main improvements made in our GIR system, namely:

Query Processing: We now handle placenames as both geographic criteria and as plain query terms. In contrast to our initial ideas, placenames revealed to be in fact good retrieval terms, and they were frequently selected as the top ranking terms in the blind relevance feedback (BRF) process [3]. While placenames may be used in other unrelated contexts, such as proper names, they seem to help retrieval recall when used as plain terms, while its geographic content can be used afterwards to refine the ranking scores and promote documents with placenames referred in a geographic context.

Text mining: We developed a new named entity recognition module, REMBRANDT, and used it as a text annotation tool to identify and classify all kinds of named entities in the CLEF collection [2]. This allows us to generate a more comprehensive geographic document signatures (D_{sig}), which is a list of geographic concepts already grounded from placenames found on each document. The D_{sig} were first introduced on last year's participation as a representation of the document's scopes, and were used to compute the geographic similarity of documents to the query's scope [3]. The D_{sig} comprise two kinds of geographic evidence: i) *explicit* geographic evidence, consisting of grounded placenames that designate geographic locations, such as countries, divisions or territories, and ii) *implicit* geographic evidence, consisting of other grounded entities that do not designate explicitly geographic locations but are strongly related to a geographic location, such as monuments, buildings, company headquarters or summits.

Document Processing: To cope with the new approaches on query processing, we needed a simple ranking model that elegantly combined the text and geographic subspace models, eliminating the need for merging text and geographic ranking scores, while still allowing us to assign a weight for each model on the retrieval. Therefore, we extended MG4J to suit our requirements for this year's experiments [12], and we chose the BM25 weighting scheme to compute a single ranking score for documents [9], using three index fields: `text` field, for standard term indexes, `explicit local` field, for geographic terms considered as explicit geographic evidence, and `implicit local` field, for geographic terms associated to the implicit geographic evidence.

The rest of the paper is organised as follows. Section 2 outlines our GIR prototype and describes in detail each module. Section 3 presents the optimising steps and the configurations selected for the submitted runs. Section 4 dissects both the official results in GeoCLEF and our post-hoc evaluation results, and Section 5 concludes the paper with insights drawn from this participation.

2 System Description

Figure 1 describes the architecture of our GIR prototype. In a nutshell, the CLEF topics are pre-processed by QuerCol, generating query strings in MG4J syntax. The CLEF documents are geo-parsed by REMBRANDT, a named entity recognition module, that plays the role as a text annotation tool and identifies

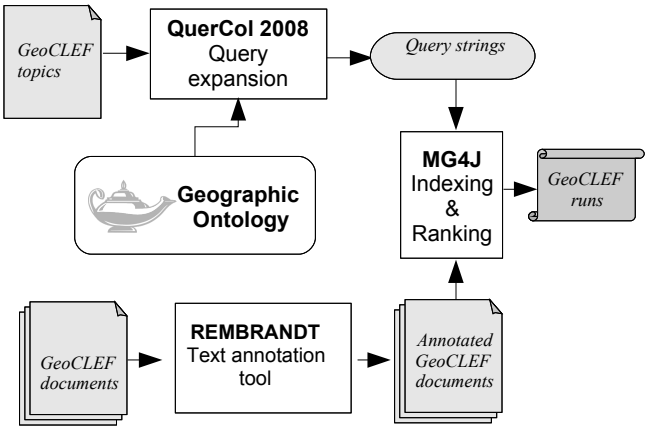

Figure 1: Architecture of the GIR prototype used in GeoCLEF 2008.

named entities that have geographic evidence, generating the geographic document signatures (D_{sig}). Afterwards, the text and D_{sig} of the documents are indexed by MG4J. The document retrieval uses an optimised BM25 weighting scheme and receives the query strings from QuerCol, generating results in the `trec_eval` format. The geographic ontology assists only QuerCol in its geographic term expansions, as REMBRANDT and MG4J use other geographic knowledge resources, as described further in this section.

2.1 REMBRANDT

REMBRANDT is a language-dependent named-entity recognition (NER) system that uses Wikipedia as a raw knowledge resource, 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 use this information to contextualise and classify surrounding named entities in the text.

One example of this additional knowledge in practice 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.

REMBRANDT currently classifies named entities using the 9 main categories and 47 sub-categories defined by the second edition of HAREM, a NER system evaluation contest for Portuguese [11, 10]. The main categories are: `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.

The REMBRANDT classification strategy relies on mapping each named entity to a Wikipedia page and subsequently analysing its document structure, links and categories, searching for suggestive evidences. REMBRANDT also relies on manually crafted rules for capturing internal and external evidence of named entities for both Portuguese and English texts, as suggested by McDonald [7]. These rules are used to classify named entities that were not mapped to a Wikipedia page or mapped to a page with insufficient information, and to contextualise named entities that have a different meaning (for example, in "I live in Portugal street", where the named entity "Portugal" designates a street, not a country).

The classification is best illustrated by following how the example named entity, "Empire State Building", is handled: the english Wikipedia page of the Empire State Building (`en.wikipedia.org/wiki/Empire_State_Building`) is labelled with 10 categories, such as "Skyscrapers in New York City" and "Office buildings in the United States". With this information, REMBRANDT classifies the named entity as a `PLACE/HUMAN/CONSTRUCTION`. In the hypothetical case that this named entity could not be mapped to a Wikipedia page, internal evidence rules, such as the presence of the term "Building" in the end, can classify the named entity as a `PLACE/HUMAN/CONSTRUCTION`. Finally, external evidence rules check the context on

Explicit geographic evidence	No geographic evidence		
PLACE/PHYSICAL/ISLAND	THING/CLASS	ABSTRACTION/DISCIPLINE	
PLACE/PHYSICAL/WATERCOURSE	THING/CLASSMEMBER	ABSTRACTION/STATE	
PLACE/PHYSICAL/WATERMASS	THING/OBJECT	ABSTRACTION/IDEA	
PLACE/PHYSICAL/MOUNTAIN	THING/SUBSTANCE	ABSTRACTION/NAME	
PLACE/PHYSICAL/REGION			
PLACE/PHYSICAL/PLANET	PLACE/VIRTUAL/MEDIA	MASTERPIECE/WORKOFART	
	PLACE/VIRTUAL/ARTICLE	MASTERPIECE/REPRODUCED	
PLACE/HUMAN/REGION	PLACE/VIRTUAL/SITE	MASTERPIECE/PLAN	
PLACE/HUMAN/DIVISION			
PLACE/HUMAN/STREET	PERSON/POSITION	TIME/GENERIC	
PLACE/HUMAN/COUNTRY	PERSON/INDIVIDUAL	TIME/DURATION	
	PERSON/INDIV.GROUP	TIME/FREQUENCY	
Implicit geographic evidence	PERSON/POSIT.GROUP	TIME/HOUR	
	PERSON/MEMBER	TIME/INTERVAL	
EVENT/PASTEVENT	PERSON/MEMBERGROUP	TIME/DATE	
EVENT/ORGANIZED	PERSON/PEOPLE		
EVENT/HAPPENING			
	VALUE/CURRENCY		
PLACE/HUMAN/CONSTRUCTION	VALUE/CLASSIFICATION		
	VALUE/QUANTITY		
ORGANIZATION/ADMINISTRATION			
ORGANIZATION/INSTITUTION			
ORGANIZATION/COMPANY			

Table 1: List of classification of NE categories and sub-categories, as having explicit, implicit or no geographic evidence for the generation of D_{sig}.

which the named entity is inserted, ensuring that the named entity is not referred in another context (for example, as an hypothetical movie, street or restaurant name). For the detection of implicit geographic evidence, the categories "Skyscrapers in New York City" and "Office buildings in the United States" are handled by REMBRANDT as additional text, and the place names "New York City" and "United States" are captured and listed as implicit geographic evidence associated to the named entity "Empire State Building".

From REMBRANDT annotations to geographic document signatures

Each CLEF document annotated with REMBRANDT contains a list of named entities that might convey explicit or implicit geographic evidence. We can now generate rich geographic document signatures D_{sig} by adding named entities that have explicit geographic evidence, together with the placenames that were associated as implicit geographic evidence for other named entities. We divide the 47 sub-categories of named entities into 3 levels of eligibility, as depicted in Table 1:

1. **Sub-categories that have explicit geographic evidence**: all sub-categories under the main category `PLACE`, with the exception of the sub-categories `PLACE/HUMAN/CONSTRUCTION` and `PLACE/VIRTUAL/*`. The category `PLACE` mostly spans the administrative domain and physical domain, but the `PLACE/VIRTUAL/*` sub-categories span virtual places such as web sites, newspaper articles or TV programs, and therefore are not eligible for inclusion in the geographic signatures. In HAREM, the subcategory `PLACE/HUMAN/CONSTRUCTION` is included in the `PLACE` main category, precisely because of its strong geographic connotation, but it is not an explicit geographic entity. As such, the subcategory `PLACE/HUMAN/CONSTRUCTION` is handled in the next level.

2. **Sub-categories that have implicit geographic evidence**: the categories `ORGANIZATION`, `EVENTS` and sub-category `PLACE/HUMAN/CONSTRUCTION`. The category `ORGANIZATION` spans institutions and corporations, such as city halls, schools or companies, which are normally related to a defined geographic area of interest. The category `EVENTS` spans organised events that normally take place in a defined place, such as olympic games, rock concerts or conferences.

3. **Sub-categories that have no geographic evidence**: categories considered to have no significative contribution for the geographic signatures. It spans the categories `PERSON`, `THING`, `ABSTRACTION`, `MASTERPIECE`, `TIME` and `VALUE`, and sub-categories `PLACE/VIRTUAL/*`.

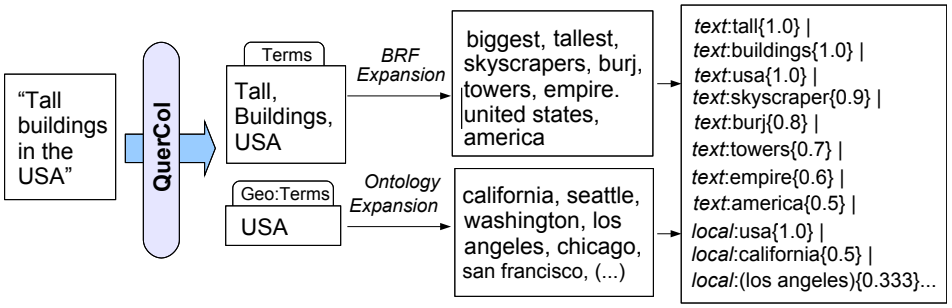

Figure 2: QuerCol's query reformulation strategy for 2008.

This eligibility table of named entity classifications into D_sig signatures is a simplification exercise, and it is far from consensual. It is questionable whether categories such as `PERSON` can also convey a significative geographic evidence to define the document scopes. For instance, the named-entity "Nelson Mandela", as processed by REMBRANDT, is associated to "South Africa" as its implicit geographic evidence because the Wikipedia page of Nelson Mandela (`en.wikipedia.org/wiki/Nelson_Mandela`) contains the category "Presidents of South Africa". Yet, this geographic evidence may cause the drift from the document scope, because not all documents mentioning "Nelson Mandela" have the South African territory as their geographic scope.

On the other hand, we are assuming that all captured geographic evidence is relevant for the document scope, but this is not always true. Take for instance the named entity example "Empire State Building"; while it conveys an implicit location when it is addressed, for example, in a context of office headquarters, it is not important for the document scope when it is addressed on a context of its architectural style.

2.2 QuerCol

QuerCol's query reformulation has two different procedures: first, it uses blind relevance feedback (BRF) for selected terms, and secondly, it performs geographic query expansion for geographic terms, by exploring the relationships between geographic concepts on a geographic ontology [5].

Figure 2 illustrates the two different expansion procedures of QuerCol, for the example query "Tall buildings in the USA". First, QuerCol removes the stopwords from the query, and recognises the geographic terms with the help of REMBRANDT. Afterwards, the non-stopwords *tall*, *buildings* and *usa* are expanded through BRF, using the $w_t(p_t\text{-}q_t)$ algorithm to weight terms in a normalised scale of [0,1]. [6] The expanded terms are then merged with the initial query terms with an OR logic operator (|), labelled with their targeted index field and corresponding term weights, and finally assembled in a single query string.

On the other hand, the geographic term "USA" is grounded to the geographic concept 'United States of America (country)', triggering the ontology-driven geographic query expansion that searches for other geographic concepts known to be contained within the USA territory. The new geographic terms are then re-weighted according to the ontology node distance between the root node and the leaf node by the formula $\frac{1}{n-1}$. For the given example, USA generates 50 states with a weight of $\frac{1}{2}$ and several cities with weight $\frac{1}{3}$ (considering that the node distance in the ontology between states and countries is 1, and between cities and countries is 2). In the end, terms are labelled as search terms for the `explicit local` and `implicit local` index fields. These list of terms is desigated as the query geographic signature, the Q_{sig}.

2.3 MG4J indexing and ranking

MG4J is responsible for the indexing and retrieval of documents. MG4J indexes the text of CLEF documents into a `text` index field, while the D_{sig} of the documents is divided in two geographic indexes: the `explicit local` and `implicit local` index fields, according to each type of geographic evidence. Figure 3 presents an example of REMBRANDT's annotation and subsequent MG4J indexing steps.

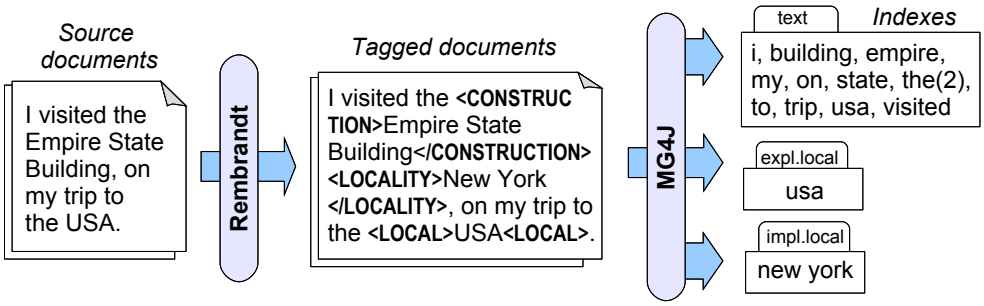

Figure 3: REMBRANDT's text annotation and MG4J's indexing steps.

We define *term similarity* as the similarity between query subjects and document subjects, and computed with the use of BM25 on the `text` index field only, and *geographic similarity* as the similarity between geographic signatures of queries and documents (Q_{sig} and D_{sig}), computed with the use of BM25 on the `explicit local` and `implicit local` index fields. MG4J allows us to dynamically select the indexes to be used in the retrieval, and to change the weight of a field before retrieval.

Unfortunately, we were not able to extend the BM25 implementation on MG4J to support term weight, so all the terms weights were set to the defalut value of 1 for all the generated runs.

3 Run Generation

The run generation procedure is depicted in Figure 4. In an initial step, QuerCol processes the topics and performs only a geographic query expansion, generating an ontology expanded query. The ontology expanded query is submitted to MG4J, generating the *initial run*. The initial run with the best men average precision (MAP) value is chosen as the source of relevance for the blind relevance feedback step, performed by QuerCol on the ontology expanded query. In the end, QuerCol generates a final BRF + ontology expanded query, which is submitted to MG4J to generate the *final run*.

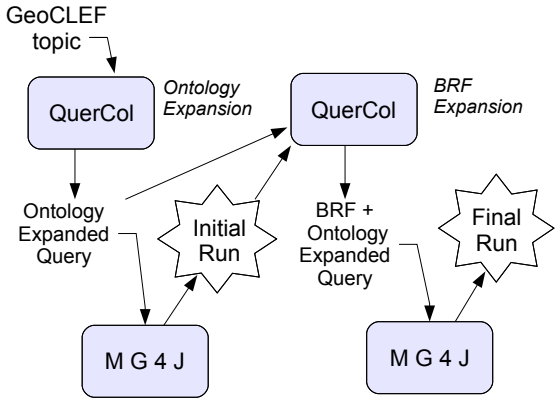

Figure 4: Run generation procedure.

3.1 Optimising the GIR parameters

We thoroughly optimised the parameters of our GIR prototype, so that we could minimise the effect of a detuned GIR system on the MAP values, and increase our confidence that the results are a direct consequence of the approaches being evaluated. Regarding the document retrieval, we optimised the b and k_1 parameters of BM25, and we experimented different weights for the `text`, `explicit local` and `implicit`

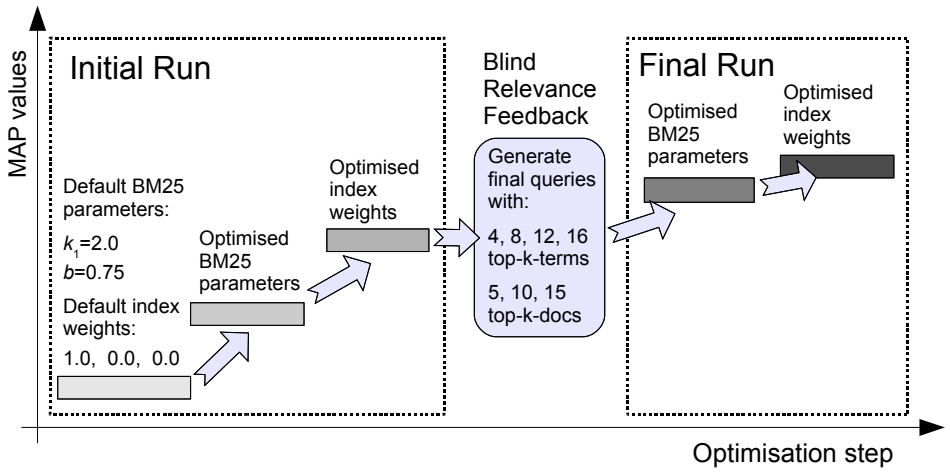

Figure 5: Optimisation procedure for the GIR system.

`local` index fields. For the QE step, we experimented with different blind relevance feedback parameters: the number of terms added for the final query, *top-k-terms*, and the number of top documents considered relevant, *top-k-docs*.

Figure 5 describes the optimisation procedure performed on our GIR prototype with the help of the 2007 GeoCLEF data. We started with the default values for the BM25 parameters, and using only the `text` index field, we searched for the best BM25 values that generated the optimal MAP values. Then, with these BM25 parameters, we then optimised the index weight values according to the MAP values. The best initial run fed the blind relevance feedback process, and the optimisation procedure was repeated for the final run stage, using several combinations of top-k-terms and top-k-docs.

3.2 Configuration of the official runs

	Initial Run					BRF		Final Run				
Run number	BM25 opt.		Index weight optimisation			top-k	top-k	BM25 opt.		Index weight optimisation.		
Portuguese	b	k_1	text	exp.l.	imp.l.	terms	docs	b	k_1	text	exp.l.	imp.l.
#1, #2, #3	0.4	0.9	{2.0, 2.5, 3.0}	0.25	0.0	-	-	-	-	-	-	-
#4, #5, #6	0.4	0.9	2.5	0.25	0.0	8	5	0.95	0.3	{2.0, 2.5, 3.0}	0.25	0.0
#7, #8, #9	0.4	0.9	2.5	0.25	0.0	8	5	0.65	0.35	{2.0, 2.5, 3.0}	0.25	0.0
#10,#11, #12	0.4	0.9	2.5	0.25	0.0	8	5	0.65	0.5	{2.0, 2.5, 3.0}	0.25	0.0
English												
#1, #2, #3	0.65	1.4	{1.5, 2.0, 2.5}	0.5	0.0	-	-	-	-	-	-	-
#4, #5, #6	0.65	1.4	2.0	0.5	0.0	8	15	0.65	1.4	{1.5, 2.0, 2.5}	0.5	0.0
#7, #8, #9	0.4	0.9	2.5	0.25	0.0	8	10	0.65	0.35	{1.5, 2.0, 2.5}	0.5	0.0
#10,#11,#12	0.4	0.9	2.5	0.25	0.0	8	5	0.65	0.5	{1.5, 2.0, 2.5}	0.5	0.0

Table 2: The configuration parameters used for the official runs.

In 2008, GeoCLEF allowed up to 12 official runs to be submitted, for each of the monolingual subtask. We submitted a total of 12 runs for each subtask, using the most promising parameters from the optimisation procedure, with a sligh variation on the index weights. Table 2 resumes the parameter values used for the official runs.

The official runs are composed by initial runs (#1 to #3) and final runs (#4 to #12). We experimented different ratios of `text` / `explicit local` index weights, by increasing and decreasing the `text` index weight by 0.5.

During the optimisation, we observed that the BM25 optimisation for the Portuguese subtask presentes many local optimal MAP values, so we decided to submit runs with three BM25 configurations from different areas, to increase the odds of standing near a global optimal BM25 parameter. For the English optimisation, we observed that the BRF parameters had more influence on the optimal MAP values than the BM25 parameters, so we submitted runs with different BRF parameter values. Also worth mentioning is the fact that the `implicit local` index field did not improved MAP values in any optimisation scenario, and thus it was turned off on all official runs.

4 Results

Best GeoCLEF 2008 runs

	Initial Run							BRF		Final Run						
	BM25 optim.			Index weight optim.				top-k	top-k	BM25 optim.			Index weight optim.			
	b	k_1	MAP	text	exp.l.	imp.l.	MAP	terms	docs	b	k_1	MAP	text	exp.l.	imp.l.	MAP
PT3	0.4	0.9	0.2222	2.5	0.25	0.0	**0.2234**	-	-	-	-	-	-	-	-	-
EN6	0.65	1,4	0.2519	2.0	0.5	0.0	0.2332	8	15	0.65	1.4	-	2.5	0.5	0.0	**0.2755**
PT	Best Optimisation values															
2006	0.4	0.4	0.1613	2.0	0.25	0.0	0.1810	16	5	0.55	0.9	0.1967	2.0	1.25	0.5	**0.2082**
2007	0.4	0.9	0.273	2.5	0.25	0.0	0.3037	8	5	0.3	0.95	**0.3310**	2.5	0.25	0.0	**0.3310**
2008	0.35	1.2	0.2233	4.0	0.25	0.0	**0.2301**	12	15	0.5	1.0	0.2069	1.5	0.25	0.0	0.2089
EN																
2006	0.3	1.6	0.2158	2.25	0.5	0.25	0.2442	16	5	0.8	0.2	0.2704	0.75	0.25	0.5	**0.2714**
2007	0.65	1.4	0.2238	2.0	0.5	0.0	0.2713	8	15	0.65	1.4	**0.2758**	2.0	0.5	0.0	**0.2758**
2008	0.65	1.6	0.2528	3.5	0.25	0.25	0.2641	12	10	0.75	0.6	0.2809	2.0	0.25	0.0	**0.2814**

Table 3: MAP values and optimising parameters for the 2008 official runs and for the optimisation step.

Table 3 presents the best GeoCLEF official runs (top part) and the best optimisation values for the GeoCLEF evaluation data from 2006 to 2008 (top part), for the Portuguese and English monolingual subtasks. We observe that our best Portuguese run was in fact an initial run (with a MAP of 0.2234), and the post-hoc optimisation corroborated the fact that the best MAP values for Portuguese are achieved by initial runs (with the best MAP value of 0.2301), which is somewhat unexpected. For the English subtask, the best run was indeed a final run, achieving a MAP value of 0.2755, that could be pushed further up to 0.2814 with optimised parameters.

The results show that the use of `explicit local` index field on the retrieval process improves the results in all evaluation scenarios, while the `explicit local` index field does not contribute at all to the improvement of the retrieval results. This fact proves that the GIR prototype is able to outperform a classic IR system in a consistently way, but it also contradicts our initial beliefs that implicit geographic evidence would have an important role on the D_{sig}. In fact, we only observe that the `explicit local` index field takes part on the best MAP values for GeoCLEF 2006, which we think that it might be related to the geographically generic topics used in that year (mostly about countries and continents), which favours the implicit geographic evidence (that is also normally given by countries and continents).

Another topic of interest of the results is that we were not so far away from the optimal MAP values as we initially expected to be, and thus we believe that we did not over-fitted our GIR system with the 2007 data. Nonetheless, the post-hoc optimisation revealed that the English topic sets are quite balanced, where we consistently achieved MAP values around 0.28, while the difficulty of the Portuguese topic sets is more unpredictable. This reveals how important it is to tune up a system according to the collections and topics, as the default parameter values rarely produce good results.

5 Conclusions

This year, we participated in GeoCLEF with the purpose of maturing the ideas first coined on last year's participation, namely: i) the document geographic signatures must be more comprehensive, extraction all kinds of geographic evidence that can be derived from all named entities in the text, and ii) the thematic and geographic facets of each search are not antagonic, and could be used together to retrieve documents in a common weighting scheme that can elegantly combine term and geographic index fields.

The results showed that our GIR prototype is consistently better when using the geographic indexes on the retrieval, meaning that our GIR approach outperforms a classic IR retrieval in every GeoCLEF evaluation scenario since 2006. For future work, we plan to improve REMBRANDT's strategy for capturing implicit geographic evidence, as we believe that its naïve approach generated noisy signatures and it was responsible for the futility of the `implicit local` index field. We also want to develop a new adaptive strategy for QuerCol, as the optimal QE parameters vary for each topic, and using the same configuration set for all topics generates sub-optimal expanded queries. We also plan to rebuild the BM25 implementation of MG4J, so that term weights can be properly used on document retrieval.

Acknowledgements

We thank David Cruz and Sebastiano Vigna for the modifications made to MG4J according to the experiments, and to Marcirio Chaves for updating the geographic ontology. Our participation was jointly funded by the Portuguese government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC (Linguateca), and partially supported by grants SFRH/BD/29817/2006 and POSI/SRI/40193/2001 (GREASE) from FCT (Portugal), co-financed by POSI.

References

[1] Guoray Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In *Proceedings of the 2nd International Conference on Geographic Information Science, GIScience'02*, pages 65–79, London, UK, 2002. Springer-Verlag.

[2] Nuno Cardoso. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In Cristina Mota and Diana Santos, editors, *Encontro do Segundo HAREM*, Aveiro, Portugal, 11th September 2008. in Portuguese.

[3] Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 802–810. Springer, 2008. Revised Selected papers.

[4] Nuno Cardoso and Diana Santos. To separate or not to separate: reflections about GIR practice. In *1st Workshop on Novel Methodologies for Evaluation in Information Retrieval, NMEIR 2008 (ECIR'2008 Workshop)*, Glasgow, UK, 30 March 2008.

[5] Nuno Cardoso and Mário J. Silva. Query Expansion through Geographical Feature Types. In *4th Workshop on Geographic Information Retrieval, GIR'07 (CIKM'2007 Workshop)*, Lisbon, Portugal, 9th November 2007. ACM.

[6] Efthimis N. Efthimiadis. A user-centered evaluation of ranking algorithms for interactive query expansion. In *Proceedings of ACM SIGIR '93*, pages 146–159, 1993.

[7] D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In I. Boguraev and J. Pustejovsky, editors, *Corpus processing for lexical acquisition*, chapter 2, pages 21–39. MIT Press, Cambridge, MA, USA, 1996.

[8] M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR'1998*, pages 206–214, Melbourne, Australia, 1998. ACM Press.

[9] Stephen E Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, pages 21–30, Gaithersburg, MD, USA, 1992.

[10] Diana Santos, Paula Carvalho, Hugo Oliveira, and Cláudia Freitas. Second HAREM: new challenges and old wisdom. In *International Conference on Computational Processing of Portuguese Language, PROPOR'2008*, Aveiro, Portugal, 8-10th September 2008. Accepted for publication.

[11] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik, and Daniel Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'2006*, pages 1986–1991, Genoa, Italy, 22-28 May 2006.

[12] Sebastiano Vigna and Paolo Boldi. MG4J: Managing Gigabytes for Java™. `http://mg4j.dsi.unimi.it/`, December 2007.