University of Twente at GeoCLEF 2006: geofiltered document retrieval

Claudia Hauff * Dolf Trieschnigg * Henning Rode †
University of Twente
Enschede, The Netherlands
{hauffc,trieschn,rodeh}@ewi.utwente.nl

Abstract

In this report we describe the approach of the University of Twente to the 2006 GeoCLEF task. It is based on retrieval by content and the subsequent filtering by geographical relevance utilizing a gazetteer. The results do not show an improvement in retrieval performance when taking geographical information into account.

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; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

1 Introduction

GeoCLEF is a track of the Cross Language Evaluation Forum (CLEF) which evaluates the retrieval of multilingual documents with an emphasis on geographic search [2]. Given a number of topics in different languages the systems have to find relevant documents in a predetermined document collection. This year's evaluation provides 25 topics which describe information needs with particular geographical references. These references vary from explicit location names such as "Car bombings near *Madrid*" to vague descriptions of geographical areas like "Wine regions around *rivers in Europe*".

In this report we describe the approach of the University of Twente to the 2006 GeoCLEF task which is based on the hypothesis that a detailed geographical thesaurus improves retrieval performance. This is our first attempt in building a Geographic Information Retrieval system and a large effort went into constructing a geographic thesaurus.

The outline of this report is as follows. In Section 2 some related work is discussed. Our approach, including the construction process of the thesaurus, is discussed in Section 3. Section 4 outlines the experiments carried out and their results. Finally Section 5 discusses the results and provides an outlook into future work.

*Human Media Interaction group
†Database group

2 GeoCLEF 2005

GeoCLEF was held as a pilot track in 2005 and is a regular track for the 2006 forum. In the 2005 track overview [2] it is noted that the best performance of that year was achieved using standard keyword search techniques ignoring the geographic references. Gey and Petras [3] reported on deteriorated performance when applying manual query expansion of geographic references. Guillén [4] concluded that including geographic information in the queries could not significantly improve retrieval performance. Metacarta [5]'s approach using geographic bounding boxes does outperform their keyword-only approach; however it's mean average precision is not higher than 17%.

3 Approach

Despite the disappointing results of last year's efforts to incorporate some spatial awareness in IR systems, we believe that adding knowledge about locality can improve the search performance. We have confined ourselves to the monolingual task and thus have only worked with the English topics and documents.

Our approach can be summarized as follows:

1. Carry out document retrieval to find "topically relevant" documents. For example, for the topic "Car bombings near Madrid" this step should result in a ranked list of documents discussing car bombings, not necessarily near Madrid.
2. Filter this ranked list based on "geographical relevance". For each topically relevant document, determine whether it is also geographically relevant. If not, it is removed from the list.

In the following sections, this approach will be further discussed. Section 3.1 discusses the construction process of the gazetteer, which is required for determining geographical relevance. Sections 3.2 and 3.3 discuss the preprocessing steps applied to the corpus and queries respectively. Section 3.4 describes the document retrieval step and in Section 3.5 the geographical filtering process is outlined.

3.1 The Gazetteer

In order to perform the geographical filtering step, each document in the document collection is tagged beforehand with appropriate geographical labels. This geotagging process requires a gazetteer which lists geographical references, links them to geographical locations (longitude and latitude values) and provides information about parent-child relationships between these references such as "*Madrid* lies in *Spain* which is part of *Europe*".

The construction of the gazetteer proved to be difficult as we relied on freely available resources and had to combine several of them to achieve the intended coverage.

3.1.1 Sources

Our merged gazetteer (MG) was derived from three freely available gazetteers:

- GEOnet Names Server (GNS)[1]
  with approximately 5.6 million entries covering the world excluding the USA and Antarctica;
- the Geographic Names Information System (GNIS)[2]
  with 1.8 million entries about the USA and associated territories and
- the World Gazetteer (WG)[3]
  with 146000 entries from all over the world.

[1] http://earth-info.nga.mil/gns/html/
[2] http://geonames.usgs.gov/stategaz/
[3] http://www.world-gazetteer.com/

	#occurrences	GNS	GNIS	WG
<entity>				
<name>Zwonitz</name>	1	x	x	x
<altname>Zwönitz</altname>	0-n	x	x	x
<latitude>50.6333333</latitude>	1	x	x	x
<longitude>12.8</longitude>	1	x	x	x
<country>Germany</country>	0-1			x
<region>Western Europe/Americas</region>	1	x	x	
<state> </state>	0-1		x	
<parent1>Sachsen</parent1>	0-1			x
<parent2>Chemnitz</parent2>	0-1			x
<parent3>Stollberg</parent3>	0-1			x
</entity>				

Table 1: Gazetteer information by combining GNS, GNIS and WG

GNIS and GNS cover most parts of the USA and Western Europe respectively very densely, which was thought to be an advantage, as the GeoCLEF corpus consists of articles from the LA Times (USA) and the Glasgow Herald (United Kingdom) - two newspapers that offer local as well as international news. GNIS and GNS only contain a small number of parent-child relationships though. Parent information (country, region, etc) does exist in the much less detailed WG and for that reason it was chosen to augment the other two gazetteers: for each entry of the WG its name and longitude/latitude values were compared against the GNIS/GNS data and if an agreement was found (matching name; longitude/latitude pair does not deviate by more then 0.05 degrees) the parent information was added.

GNS and GNIS were also utilized by MIRACLE [6] in last year's task. A combination of all three gazetteers was employed by the GeoTALP system [1], however no additional parent information was inferred.

A typical entry of MG is shown in Table 1. Each tag is listed with the possible number of occurrences per entry and the source gazetteer(s). A location (latitude, longitude pair) may be known under several names (different spellings, short forms) and these possibilities are listed under the tag alternative names. For entries covering the USA, the state is also given. The parent-tags have different granularities - parent1 is the most general and parent3 the most specific.

3.1.2 Preprocessing and Statistics

The coverage of MG is not uniform: whereas the USA and Western Europe are well represented, other regions - such as Canada, Northern Africa, a large part of Asia - are barely covered. In order to acquire a better understanding of MG's density distribution of geographic locations, the density was plotted on a grid (Figure 1). Grid regions with few gazetteer entries are green, while red areas are densely covered.

Not all entries in MG contain data for all tags. Only name, latitude, longitude and region are guaranteed to exist for each entry. As can be seen from the number of tag occurrences in MG (Table 2), especially parent information is scarce. This is due to the parent-child relationship information coming from the relatively small WG. A simple algorithm was applied to infer more relationships for nearby locations, which are not directly covered by the WG.

In a first step, all entries with parent-child information were sorted in a grid representing the world map. For each entry without parent information the appropriate cell in the grid was determined. All entries with a parentX ($X = 1, 2, 3$) in the same or adjacent (north, south, east or west) cells were utilized for inferring new parentX information for the entry. We will refer to these entries as *inferring entries*. Two strategies were tested: full agreement and the less restrictive majority agreement:

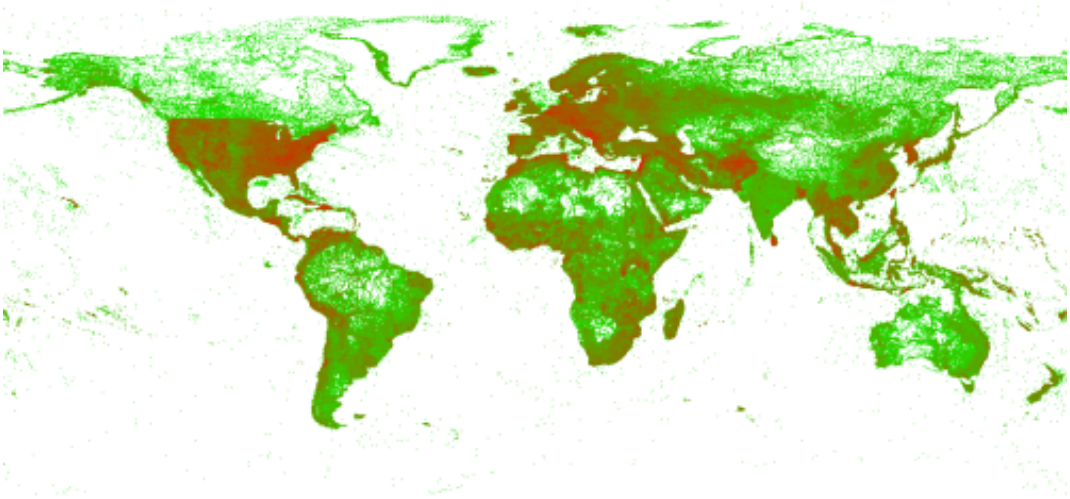

Figure 1: Location density of the merged gazetteer.

type	#names
name (incl. altname)	9019583
country	142086
region	5327830
state	1761884
parent1	142085
parent2	56448
parent3	17864

Table 2: Merged gazetteer tag occurrences.

full agreement all inferring entries have to agree on a common parentX element in order to assign it to the entry; furthermore at least two entries with parentX elements need to exist, otherwise parent information is not assigned.

majority agreement a majority of the inferring entries have to agree on a parentX; for the parent2 and parent3 assignment, only those entries are utilized that contain the 'winning' parent1 or parent2 element respectively.

The grid resolution was varied between 1 and $0.16\bar{6}$ square degrees. For the latter resolution, the number of inferred parents via majority and full agreement voting are listed separately for the USA and the World (excluding the USA) in Table 3. The parent information inferred from the majority voting were utilized and inserted into the MG. Using majority votes over full agreement can be justified by the high resolution.

3.1.3 Unsolved problems

During the course of the experiments we came across several problems, that have not been adequately solved yet.

- Due to MG's detailed coverage, some names are highly ambiguous and appear over a thousand times as different entries. The ten most ambiguous names are listed in Table 4.
- The GNS/GNIS data also contains entries of geographic entities that stretch over a certain area (like rivers or large cities), but only a single latitude/longitude pair is provided for them.

type	USA	World
p1 full	1141146	1289983
p2 full	0	107760
p3 full	0	5622
p1 maj	1386462	2008810
p2 maj	0	359876
p3 maj	0	227830

Table 3: Number of inferred parents on full and majority agreement voting.

name	#occurrences
first baptist church	2020
san jose	1736
the church of jesus christ of latter day saints	1721
san antonio	1713
san josé	1483
mill creek	1472
spring creek	1431
church of christ	1383
santa maria	1259
dry creek	1239

Table 4: The MG's ten most ambiguous names.

- The gazetteer data is noisy. During the parsing process a number of out of range longitude/latitude pairs were encountered. Furthermore, the parent information can overlap or is not fully accurate. In Table 1 for example Zwonitz is indeed a town in the district Stollberg, Chemnitz however is not a parent of Stollberg, but a neighboring city.
- The merging process also led to inconsistencies, as the same parents or regions could have been assigned different names in the three gazetteers, for example 'USA' versus 'United States of America'.

3.2 Geotagging the Corpus

The GeoCLEF corpus consists of newspaper articles from the Glasgow Herald (GH) and the LA Times (LA).

In order to determine the geographic range of an article, potential location phrases need to be identified in the document text. One of the difficulties here are partial matches that lead to false positives when applying simple string matching. For example the phrase 'George Washington' might be falsely recognized as the location 'Washington'. This can be overcome by searching for the longest phrase of capitalized letter strings (stopping at punctuation) and matching the whole phrase against the gazetteer. While this solves the 'George Washington' problem, compound words containing lowercase conjunctions such as 'Statue of Liberty' are falsely identified as two potential locations: 'Statue' and 'Liberty'. For this reason, the capitalized phrase rule was amended: if two capitalized strings are connected by the conjunction 'of' they are considered as 1 potential location. Once a list of potential locations is extracted for each document, it is matched against the (alternative) names in MG. That this potential location matching method has its difficulties becomes apparent when looking among others at the phrase 'United Kingdom'. In the corpus this phrase always refers to 'Great Britain and Northern Ireland'; however, there exist a number of kingdoms throughout the world and therefore in the gazetteer this region is listed under the name 'United Kingdom of Great Britain and Northern Ireland'.

For each corpus document, the detected geographical references were recorded in a database.

3.3 Preprocessing the Queries

Contrary to last year, the queries are not geographically tagged. We did this manually for the topic title by tagging the location name and the type of spatial relation (around, north, east, etcetera). The topic descriptions and narratives were not tagged.

We identified six different query categories:

- **Queries with concrete locations** "Diamond trade in Angola and South Africa" (GC029), "ETA in France" (GC049)
- **Queries with locations and simple rules of relevant locations** "Cities within 100km of Frankfurt" (GC027), "Car bombings near Madrid (GC030)
- **Queries with locations and complex rules of relevant locations** "Wine regions around rivers in Europe" (GC026), "Automotive industry around the Sea of Japan" (GC036)
- **Queries with very general locations that are not necessarily in a gazetteer** "Snowstorms in North America" (GC028), "Russian troops in the southern Caucasus" (GC039)
- **Queries with quasi-locations (e.g. political) that are not found in a gazetteer** "Malaria in the tropics" (GC034), "Credits to the former Eastern Bloc" (GC035)
- **Queries describing characteristics of the geographical location** "Cities near active volcanoes" (GC040)

A number of queries require additional world knowledge that is not covered by MG: Which entries describe rivers and volcanoes? What locations does a river flow along? Furthermore in some instances knowledge is required that cannot be found in a gazetteer: Which volcanoes are active? Which countries form the former Eastern Bloc?. Hence it is not sufficient to rely on gazetteer information alone, other sources of knowledge need to be taken into account. For our experiments we utilized Wikipedia[4] as a source of additional world knowledge.

Given a query, it's potential locations are extracted. For a tagged query this simply is the text between the location tags. For an untagged query all capitalized letter phrases (determined as described in Section 3.2) are location candidates and are matched against the gazetteer. If no match is found for any of the candidates, the extraced phrases are utilized as a Wikipedia query and the returned page is geotagged the same way as the corpus documents.

The spatial relations are only taken into account for queries where matching locations are found directly in the gazetteer (the returned Wikipedia page is too noisy). If the location entry is a country, its boundaries (minimum and maximum latitude/longitude pairs in the gazetteer) are applied as location coordinate restrictions. For the relation 'around' a coordinate restriction of ± 1 degree was used.

3.4 Document Retrieval

The corpus was indexed with the Lemur Toolkit for Language Modeling and Information Retrieval[5]. Stopwords were not removed and stemming was not performed due to the process of extracting potential location phrases from the text. The basic corpus statistics are given in Table 5.

[4] http://www.wikipedia.org
[5] http://www.lemurproject.org/

	GH	LA
number of documents	56472	113005
number of terms	24826294	63802290
number of unique terms	163252	230803
average document length	439	564

Table 5: Corpus Statistics for the Glasgow Herald (GH) and the LA Times (LA).

run id	title	desc.	narr.	geo	merged	map
baseline	x					17.45%
utGeoTIB	x			x		16.23%
utGeoTIBm	x			x	x	17.18%
baseline	x	x				15.24%
utGeoTdIB	x	x		x		7.32%
baseline	x	x	x			18.75%
utGeoTdnIB	x	x	x	x		11.34%
utGeoTdnIBm	x	x	x	x	x	16.77%

Table 6: Results for the English task of GeoCLEF 2006.

3.5 Geographical filtering

The list of ranked documents were then filtered to remove documents outside the wanted geographical scope. For each ranked document all location names were returned from the database. For each name, all possible location entries from the gazetteer were also returned (for example there are 20 'Madrid' entries in MG) and each was compared against existing location coordinate restrictions. If a location name appeared as a parent and as a child in the gazetteer, only the parent entry was considered. Furthermore, for each child entry, its parents were recorded as well. Hence, an article mentioning 'Paris' will also record 'France' as geographic entry. If one or more restrictions were in place, only those entries that do not violate them were kept. A document that was left with at least one location entry that fulfills the coordinate restrictions was deemed relevant.

For queries without coordinate restrictions, the sets of query and document locations were split into parents sets Q_p (query parents) and D_p (document parents). Here, parents are defined as location names that appear as region, state, country, parent1, parent2 or parent3. The children sets Q_c (query children) and D_c (document children) are the location names that appear in the gazetteer but not as a parent. In order to determine geographical relevance the intersection sets $I_p = Q_p \cap D_p$ and $I_c = Q_c \cap D_c$ were evaluated. If $Q_x \neq \emptyset$ with $x = \{p, c\}$, then $I_x \neq \emptyset$ had to hold in order for the document to be geographically relevant.

4 Experiments and Results

The language modeling approach with Jelineck-Mercer smoothing ($\lambda = 0.85$) was applied to retrieve the initial content-only ranking. We tested different variations of the usage of title, description and narrative as well as merging the filtered results with the content-only ranking by adding the top filtered-out results at the end of the ranking. The results are given in Table 6. The baseline run in each case is the content-only run.

In all runs, independent of the topic part (title, description, narrative) utilized as a query, the content-only baselines outperformed the geographically filtered results. Apart from the title-runs, where the differences were small, the retrieval performance decreased drastically.

5 Discussion & Conclusion

At the time of writing we have no definite explanation for the disappointing retrieval results. We suspect a bug in our geographical filtering system but other explanations for the poor results are also possible: the returned Wikipedia pages are too noisy, they contain location names in their texts that are not actually part of the sought location. The request "*North America*" for example returns a Wikipedia entry that starts with

> *North America* is a continent in the Earth's northern hemisphere and almost fully in the western hemisphere. [...] It is the third-largest continent in area, after *Asia* and *Africa*, and is fourth in population after *Asia*, *Africa*, and *Europe*.

A second source of failure can be the large size of the gazetteer. Due to the many millions of entries, location names that most humans would associate with a single location (such as that *Madrid* lies in Spain) appear as several locations all around the world. A possible solution is to assign importance scores to locations with a single name, based on the number of inhabitants for towns and cities for example. Further directions for future work are a probabilistic matching function and taking into account the relation between locations within a document.

Acknowledgements

We gratefully acknowledge the funding from the research programmes that made this work possible: the contribution by C. Hauff and H. Rode was funded by the Dutch BSIK programme MultimediaN[6]. The contribution by Dolf Trieschnigg was funded by the Dutch BSIK programme BioRange[7] .

References

[1] Daniel Ferrés, Alicia Ageno, and Horacio Rodriguez. The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis, and a Geographical Thesaurus. In *GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview*, 2005.

[2] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview, 2005.

[3] Fredric Gey and Vivien Petras. Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In *Working Notes for the CLEF 2005 Workshop*, 2005.

[4] Rocio Guilln. CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks. In *Working Notes for the CLEF 2005 Workshop*, 2005.

[5] András Kornai. MetaCarta at GeoCLEF 2005. In *GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview*, 2005.

[6] Sara Lana-Serrano, José Goñi-Menoyo, and José González-Cristóbal. MIRACLE's 2005 Approach to Geographical Information Retrieval. In *GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview*, 2005.

[6] http://www.multimedian.nl/
[7] http://www.nbic.nl/biorange/