GIR Experimentation

Andogah Geoffrey
Computational Linguistics Group
Centre for Language and Cognition Groningen (CLCG)
University of Groningen
Groningen, The Netherlands
g.andogah@rug.nl, annageof@yahoo.com

Abstract

Geographic Information Retrieval (GIR) community has generally accepted the thesis that both thematic and geographic aspect of documents may be useful for GIR. This paper describes a preliminary experiment exploring this thesis by seperately indexing/searching geographical relevant-terms (place names, geo-spatial relations, geographic concepts and geographic adjacetives) extracted from reference document collection. Two indexes were created one for extracted geographic relevant-terms (i.e. document footprint) and one for reference document collections. Geo-Score and Thematic-Score against document collection footprint and reference document collection respectively were combined through a linear interpolation to obtained the final score for document relevance ranking. We used several freely available geographic resources – Wikipedia, World-Gazetteer, GEOnet Name Server (GNS), and WordNet. Apache Lucene was used as an indexing and search platform while Alias-I LingPipe was used to detect geographic named entities (GNEs), and other geo-relevant concepts and terms in documents. We submitted runs for monolingual English task, and our system achieved mean average precision (MAP) of 0.1690 to 0.2194. No significant improvement was observed through geographic query expansion.

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 Managment**]: Languages—*Query Languages*

General Terms

system architecture, performance, experimentation

Keywords

geographic information retrieval, geographic query expansion, geographic named entity tagging, document footprints, geographic knowledge base, relevance ranking

1 Introduction

Geographic Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Geographic information pervades many documents, and therefore, geographic references may be important for Information Retrieval (IR). Additionally, many documents contain geographic references expressed in multiple languages which may or may not be the same as the query language.

To perform GIR both themetic (non-geographic aspect) and geographic aspect of documents require consideration. In order to approach this thesis we derive for each document in the collection a corresponding document footprint containing place names (e.g. Uganda), geo-spatial relations (e.g. west of), geographic concepts (e.g. country) and geographic adjectives (e.g. Ugandan). The document footprint and reference document collection provided were separately indexed and searched. Geo-Score and Thematic-Score against document collection footprint and reference document collection were combined through a linear interpolation to obtained the final score to perform document relevance ranking. Queries were performed for geographic relevant terms identified in topics against document collection footprints and reference document collection provided to investigate impart of geo-references and geo-relevant terms for GIR.

Freely available geographic resources (from: Wikipedia[1], World-Gazetteer[2], GEOnet Name Server[3] (GNS), WordNet[4]) were consulted for query geographic reference expansion. Apache Lucene[5] was used as an indexing and search platform while Alias-I LingPipe[6] was used to detect geographic named entities (GNEs) and other geo-relevant concepts and terms in documents.

2 GeoCLEF 2006

GeoCLEF evaluation track was run for the first time at CLEF 2005 to evaluate retrieval of multilingual documents with an emphasis on geographic search [Gey et al, 2005]. As GeoCLEF 2005, GeoCLEF 2006 outline the following challenges to GIR in a multilingual environment: (1) translation of locations (e.g. Uganda (EN) to Oeganda (NL)), (2) resolution of geographic reference ambiguities (e.g. "Jack *London*" the author not a place; South Yorkshire and S. Yorks refer to the same place), (3) resolution of spatial ambiguity (e.g. Sheffield in UK or USA), (4) finding or creating suitable multilingual geographic knowledge base, and (5) combining both text and spatial retrieval methods. The specific aims for GeoCLEF 2006 are: (1) compare methods of query translation, (2) query expansion, (3) translation of geographical references, (4) use of text and spatial retrieval methods separately or combined, and (5) retrieval models and indexing methods.

GeoCLEF 2006 consists of document collections in English, German, Portuguese and Spanish, and 25 search topics in these languages. The tasks for GeoCLEF 2006 are: (1) monolingual retrieval – retrieval where the topic and document languages are the same, and (2) bilingual retrieval – cross-language retrieval where the topic language is different from the document language, i.e. X → {DE, EN, ES, PT}. For each document language, participants may submit the results of up to 10 runs: 5 monolingual and 5 bilingual. Two of these runs are required: (1) Title-Description – where the search queries are created using only the contents of the Title and Desc tags of the topic, and (2) Title-Description-Narrative – where the search queries are created using the contents of the Title, Desc and Narr tags from the topic. The Narrative tag contains a more comprehensive description of the information request defined by the topic, including specifics about the geography of the topic such as a list of desired cities, states, countries, rivers or latitudes and longitudes. An example search topic is depicted below:

<top>
<num>GC027</num>
<EN-title>Cities within 100km of Frankfurt</EN-title>
<EN-desc>Documents about cities within 100 kilometers of the city of Frankfurt in Western Germany</EN-desc>
<EN-narr>Relevant documents discuss cities within 100 kilometers of Frankfurt am

[1] http://www.wikipedia.org
[2] http://www.world-gazetteer.com
[3] http://earthinfo.nga.mil/gns/html
[4] http://wordnet.princeton.edu
[5] http://jakarta.apache.org/lucene
[6] http://alias-i.com/lingpipe

Main Germany, latitude 50.11222, longitude 8.68194. To be relevant the document must describe the city or an event in that city. Stories about Frankfurt itself are not relevant</EN-narr>
</top>

3 Previous works

GeoCLEF 2005 [Gey et al, 2005] featured several approaches to GIR: (1) conventional IR systems, (2) geographic named entity recognition, classification and real world resolution, (3) creation and expansion of geographic knowledge base (e.g. name variants, multilingual), (4) query expansion strategies – blind feedback, addition of proper names, geographic reference expansion using hierarchical information contained in GKB, (5) geo-spatial query restriction strategies – minimum bounding box based, geo-scope based, and (6) topic translation strategies mainly employing usage of off-shelf software packages.

Larson [2005] provided three Lucene index types: verified place names (an index of names which matched the gazetteer entries), point coordinates (latitude and longitude coordinates of the verified place name) and bounding box coordinates (bounding boxes for the matched places from gazetteer). Text indexes were also created for separate XML elements (such as document titles or dates) as well as for the entire document. The authors found blind feedback to improve query results.

Ferres et al [2005] provided a Lucene derived Document Retrieval component which extracted relevant documents likely to contain the user information in the query. The Document Retrieval phase provides for: (1) query type (boolean query, ranked query, boolean+ranked query), (2) geographic search mode (lemma field and geo field), and (3) geographical search policy (strict search and relaxed search). Document ranking component joins the documents provided by the Document Retrieval phase.

Hughes [2005] describe loosely aggregated system for GIR comprising of gazetteer, named entity taggers and conventional IR system. The topic and document (headlines only) collections were geographical resolved by using the named entity taggers and gazetteer. This analysis allows for expansion or reduction of geospatial entities by hierarchy traversal in the gazetteer. Document collections (textual content only) were then indexed. The difference of this experiment is in the inclusion of various parts of the topics and the level of geospatial entity expansion based on the topic to geospatial entity mapping tables. However, the author found no overall performance increase by use of topics expanded with geospatial entities over the baseline topics.

Buscaldi et al [2005] describes a query expansion method based on the expansion of geographical terms by means of WordNet synonyms and meronyms. Examples of geographic synonyms: Rome (EN) and Roma (IT), U.S and U.S.A (acronyms), etc. Examples of geographic meronyms: Washington referencing U.S.A, Paris without France explicitly mentioned in the context, thus resolved to Paris, France because assumed to be well-known. The WordNet resolves synonyms through *synset* and meronyms through *part-of* relationship. The authors noted that query expansion did not provide a clear advantage and actual performed worse compared conventional search strategies. One probable reason is that the query expansion could have introduced unnecessary information. However, using WordNet synonyms and holonyms during indexing proved useful with better performance. A named entity detector was used to recognize location named entities. For every location name l, the synonyms of l and all its holonyms (e.g. Los Angeles → California → United States → North America → America) are added to the *geo* index.

Berkeley group 2 [Gey and Vivien, 2005] retrieval strategy involved query augmentation with *blind feedback*. Another feature of their approach is the augmentation of query information by inclusion of location-specific tags and expansion of geographic references (e.g. Europe to individual country names). The blind feedback approach adds 30 top-ranked terms to the query from the top 20 ranked documents of intial ranking. Manual expansion of geographic references proved disastrous to retrieval performance. Addition of *concepts* and *location* information improved retrieval precisions across. Most improvement was achieved with blind feedback by adding mostly proper names and word variations and very few irrelevant words that won't distort the search towards another direction.

4 Our appraoch

We are participating in GeoCLEF evaluation track at CLEF 2006 for the first time. The main motivation for our participation is to experiment with both thematic and geographic aspect of a document for GIR. In this section we describe our approach, system architecture and resources used. Our appraoch borrows techniques from (Larson [2005], Ferres et al [2005], Hughes [2005], Buscaldi et al [2005], Gey and Vivien [2005], Leidner [2005]) with few exceptions such as the creation of an index of document collection footprint along side the index of reference document collection, and thereby combining query results of the two index searches using linear interpolation.

4.1 Resources

4.1.1 Geographic Knowledge Base

We used the World Gazetteer, GEOnet Names Server (GNS), Wikipedia and WordNet as the bases for our Geographic Knowledge Base (GKB) for several reasons – free availability, multilingual (English, Germany, Portuguese and Spanish), most popular and major places, etc. Volcano active region, European river, Atlantic Ocean ports/coast and European Wine processing region information were specifically gathered from the Wikipedia.

4.1.2 GeoTagger

Alias-I LingPipe was used to detect named entities (location, person and organisation), geographic concepts (continent, region, country, city, town, village, etc.), spatial relations (near, in, south of, north west, etc.) and locative adjectives (e.g. Ugandan).

4.1.3 GeoCoder

We used a simple appraoch to geo-code identified geographic named entities (GNEs) presented in CLIN 2005 [Andogah, 2005]. The approach exploits location type (e.g. city, mountain) and hierarchy information integrated in GKB to ground GNEs.

4.1.4 Lucene Search Engine

Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Lucene's default similarity measure is based on vector space model[7] (VSM).

4.2 Document Pre-processing

Documents were pre-processed using the Alias-I LingPipe to detect place names (e.g. Kampala), geographic concepts (e.g. city), spatial relations (e.g. west of) and adjectives referring to things or people or language connected to a place (e.g. Ugandan).

Candidate locations for detected place names is obtained from our GKB. Place names are resolved to their respective locations using a simple geo-coding approach exploiting location type and hierarchical information present in GKB. The preliminary experimental result of geo-coding approach used here was reported in [Andogah, 2005][8]. However, due to time limitation geo-coding task was not experimented as planned, instead we assume that all geo-relevant terms detected

[7]The vector space model (VSM) is an algebraic model used for information filtering and information retrieval. It represents natural language documents in a formal manner by the use of vectors in a multi-dimensional space. *http://en.wikipedia.org/wiki/Vector_space_model*

[8]http://www.science.uva.nl/events/CLIN2005/Program/Abstracts/abstract-andogah.html

in a document will some-how relate or point to a specific geographic region/scope or geographic concept in the discourse.

4.3 Indexing document collection

Footprint document collection repository derived from document collection was created. Footprint documents contain geo-relevant terms such as place name, geographic concepts, spatial relations, locative adjectives plus their respective term frequency as depicted below.

<GeographicTermFrequency docid=”GH950102-000006”>
<GT name=”east” tf=”2” gtt=”SPR” />
<GT name=”america” tf=”1” gtt=”LOC” />
<GT name=”new york” tf=”2” gtt=”LOC” />
<GT name=”buffalo” tf=”3” gtt=”LOC” />
<GT name=”orlando magic” tf=”1” gtt=”LOC” />
<GT name=”american” tf=”4” gtt=”GAD” />
<GT name=”texas” tf=”1” gtt=”LOC” />
<GT name=”buffalo jills” tf=”1” gtt=”LOC” />
<GT name=”city” tf=”1” gtt=”GCO” />
</GeographicTermFrequency>

Derived footprint documents were indexed using Lucene along side index of reference document collection provided for the experiment (see [Table 1] for details).

Table 1: Footprint document index structure

Field	Lucene Type	Description
nm	Field.Keyword	Geo-relevant term e.g. Kampala, city, west, Ugandan
tf	Filed.Keyword	Geo-relevant term frequency
gtt	Field.Keyword	Geo-relevant term type e.g. LOC (location), GCO (concept), SPR (spatial relation), GAD (geo-adjective)
docid	Field.Keyword	Document unique identification/number

Geo-relevant term frequency is factored into the index by adding the same geo-relevant-term to the index the number of times it occurences (e.g. american in the above sample footprint document is added 4 times during indexing).

Reference document collection provided for experimentation were indexed using Lucene. Document HEADLINE and TEXT contents were combined to created document content for indexing (see [Table 2] for details).

Table 2: Reference document index structure

Field	Lucene Type	Description
docid	Field.Keyword	Document unique identification/number
content	Field.Unstored	Combination of HEADLINE and TEXT tag content

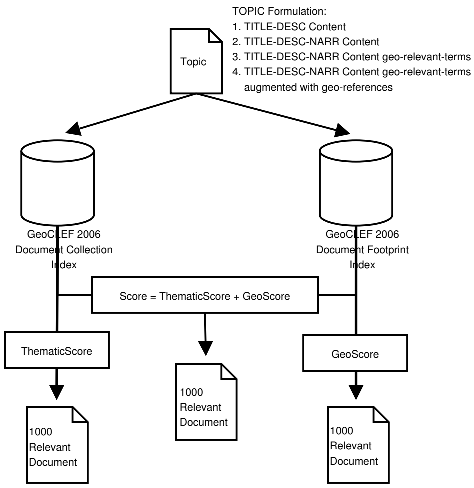

Figure 1: System architecture

4.4 Querying document collection

Mandatory runs 1 and 2 queries were formulated by topic TITLE-DESC (CLCGGeoEE1) and TITLE-DESC-NARR (CLCGGeoEE2) contents respectively. These queries were submitted to search Lucene index (Lucene field `content` was searched) of GEO-CLEF 2006 document collection (see [Table 2] for index structure and [Figure 1] for system architecture). The mandatory queries perform general-purpose search of Lucene index returning the top 1,000 documents retrieve.

Our third run query was formulated by topic TITLE-DESC only (CLCGGeoEE5). The query was submitted to search Lucene index (Lucene field `nm` was searched) of GEO-CLEF 2006 document collection footprints (see [Table 1] for index structure and [Figure 1] for system architecture), and the top 1,000 documents retrieved.

Our fourth run (CLCGGeoEE10) combine run 2 query result with result of querying Lucene index of GEO-CLEF 2006 document collection footprints for geo-relevant-terms extracted from topic TITLE-DESC-NARR. To combine the result of run 2 with result of querying Lucene index of document collection footprints we used the linear interpolation as described in [Leidner, 2005].

$$Score(d, q) = \lambda ThematicScore(d, q) + (1 - \lambda)GeoScore(d, q) \quad (1)$$

For this experimentλ was set to 0.5.

Our fifth run (CLCGGeoEE11) is similar to run four except that geo-revelant-terms extracted from topic TITLE-DESC-NARR were augmented with geo-references obtain from our GKB. For example, topic G033 geo-relevant-terms were augmented with the names of the major cities/towns/places within Ruhr area of Germany – Bochum, Bottrop, Dortmund, Duisburg, Essen, Gelsenkirchen, Hagen, Hamm, Herne, Mlheim, Oberhausen, Recklinghausen, Ennepe-Ruhr, Unna, Wesel, Mlheim an der Ruhr, Mulheim an der Ruhr.

5 Evaluation and discussion

5.1 Evaluation

Tables 2 & 3, and Figure 2 depicts the result of our official runs. The result of CLCGGeoEE5 (third run query option) is particularly interesting to note as it query GEO-CLEF 2006 document collection footprints. This scheme performed relatively well (MAP of 0.1757 compared to the best performing scheme CLCGGeoEE11 of MAP 0.2194) showing that the geographic aspect of documents are important for geographic information retrieval. Though both CLCGGeoEE1

Table 2: Average Precision at Interpolated Recall Levels

Interpolated Recall (%)	CLCGGeoEE1	CLCGGeoEE2	CLCGGeoEE5	CLCGGeoEE10	CLCGGeoEE11
0	0.3337	0.4521	0.3063	0.3792	0.4921
10	0.2845	0.3714	0.2943	0.3089	0.4024
20	0.2508	0.3061	0.2634	0.2440	0.3110
30	0.2299	0.2902	0.2406	0.2287	0.2911
40	0.2140	0.2695	0.2295	0.1997	0.2613
50	0.2037	0.2508	0.2252	0.1934	0.2461
60	0.1752	0.2040	0.1557	0.1356	0.1937
70	0.1273	0.1443	0.0964	0.0970	0.1328
80	0.0908	0.1066	0.0802	0.0875	0.1112
90	0.0721	0.0852	0.0635	0.0583	0.0897
100	0.0488	0.0631	0.0350	0.0362	0.0663

Table 3: Individual Run Performance as measured by Mean Average Precision and R-Precision

	CLCGGeoEE1	CLCGGeoEE2	CLCGGeoEE5	CLCGGeoEE10	CLCGGeoEE11
MAP	0.1730	0.2163	0.1757	0.1690	0.2194
R-Precision	0.1983	0.2194	0.1777	0.1762	0.2144

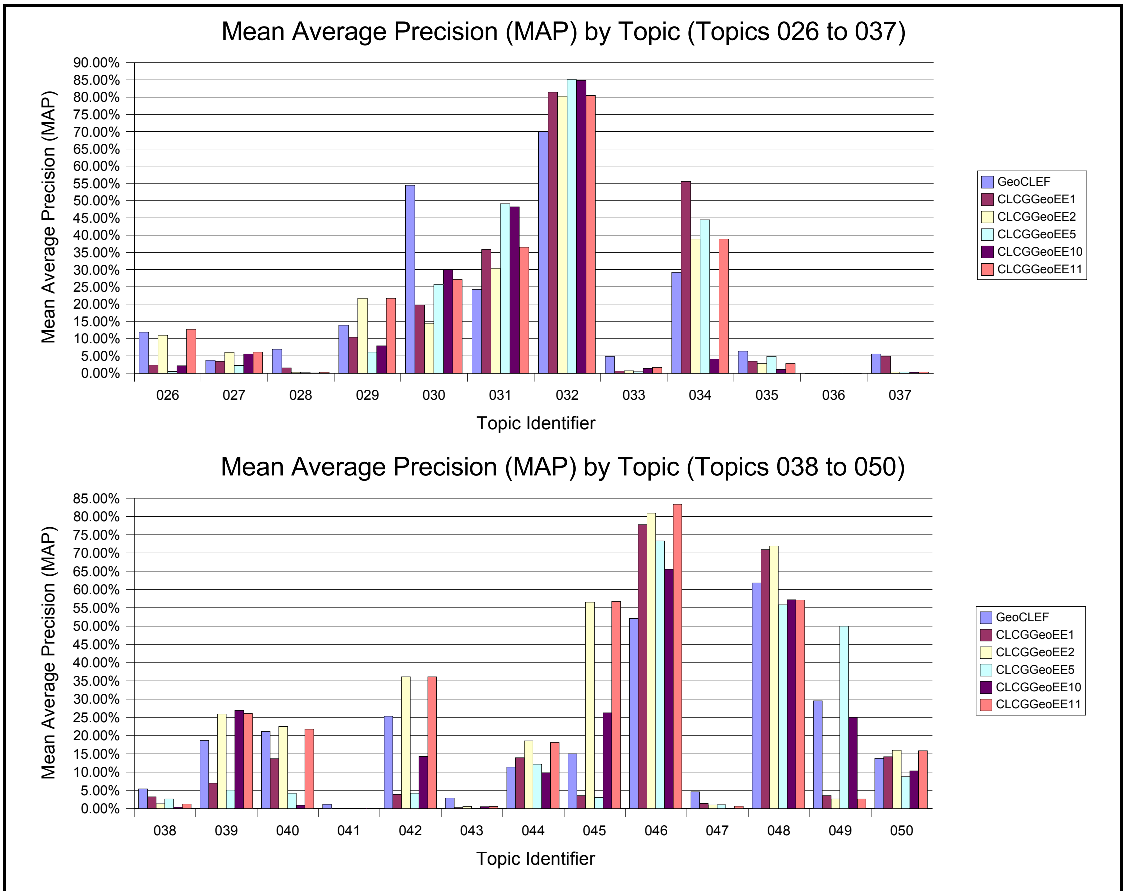

Figure 2: Mean Average Precision by Topics (Topics 026 to 050)

and CLCGGeoEE5 use topic TITLE-DESC (but querying different document collection content), CLCGGeoEE5 performed better.

CLCGGeoEE10 & CLCGGeoEE11 schemes use linear interpolation (with λ set to 0.5) to combine result of query against reference document collection and document collection footprint indexes. We note that CLCGGeoEE10 performed poorly while CLCGGeoEE11 performed better.

5.2 Discussion and future work

Several factors might have influenced the performance of these schemes (CLCGGeoEE5, CLCGGeoEE10 & CLCGGeoEE11):

- predominance of geographic concepts and spatial-relationship qualifiers such as country, city, southern, west, etc. both in the query and document footprints at expense of place names, and thereby shifting query result in wrong direction propagating irrelevant documents to the top
- value of 0.5 asigned to λ in linear interpolation [Equation (1)] above might have tilted result by asigning higher scores to documents retrieved from reference document collection or vice versa, and thereby propagating irrelevant documents to the top in the final rank
- not all documents were indexed as our adopted geographic named entity tagger (Alias-i Lingpipe) reported content error for certain files while processing reference collection files. As a result 51,525 Glasgow Heralds documents were indexed out of 56,472 and 112,552 LA Times documents were indexed out of 113,005. This might have had a considerable impart on query result as 5,400 documents (which might have contained relevant documents) were left out.

The results of our submitted runs raised several pertinent questions for future investigation:

- extend to which geographic aspect of document influence GIR result: (1) querying topic geographic aspect against reference document collection, (2) querying topic non-geographic aspect against reference document
- an appropriate value for λ in linear interpolation [Equation (1)] above for GIR
- an appropriate document collection footprint indexing strategy
- improve geographic named entity recognition, classification and real world resolution
- geographic query expansion strategies – blind feedback, addition of place names, expansion through hierarchical information contain in GKB.

6 Concluding remarks

We employed a strategy of separately indexing document footprint along side index of reference document, and combine query results of the two indexes through linear interpolation. Our approach yielded an average result as compared to overall GeoCLEF 2006 result on monolingual English task. A number of pertinent questions were raised for future investigation which we hope to address and integrate in our system. Analysis of individual topic performance to give further insight in our approach is under way.

7 Acknowledgements

This work is supported by NUFFIC within the framework of Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education and Capacity (NPT) under project titled "Building a sustainable ICT training capacity in the public universities in Uganda".

References

Gey et al [2005] Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough and Vivien. *GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Gey and Vivien [2005] Fredric Gey and Vivien Petras. *Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Larson [2005] Ray R. Larson. *Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Andogah [2005] Geoffrey Andogah. *Is Groningen referring to the city or the province? Geographic Named Entity Disambiguation.* Presentation at CLIN 2005, The 16th Meeting of Computational Linguistics in the Netherlands Amsterdam, December 16, 2005.

Leidner [2005] Jochen L. Leidner. *Preliminary Experiments with Geo-Filtering Predicates for Geographic IR.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Ferres et al [2005] Daniel Ferres, 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.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Hughes [2005] Baden Hughes. *NICTA i2d2 at GeoCLEF 2005.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.

Buscaldi et al [2005] Davide Buscaldi, Paolo Rosso, Emilio Sanchis Arnal. *A WordNet-based Query Expansion method for Geographical Information Retrieval.* At GeoCLEF 2005 in CLEF 2005 Workshop, 21 - 23 September 2005, Vienna, Austria.