A GIR Architecture with Semantic-flavored Query Reformulation

Nuno Cardoso
University of Lisbon
Faculty of Sciences, LaSIGE
ncardoso@xldb.di.fc.ul.pt

Mário J. Silva
University of Lisbon
Faculty of Sciences, LaSIGE
mjs@di.fc.ul.pt

ABSTRACT

Most geographic queries include references to entities (geographic and non-geographic). Grounding such entities is essential to properly understand the user's information need. As statistical-based query reformulation strategies work at term level, not entity level, they don't use the semantic information given by such entities, which is considerably relevant for the types of queries that should be handled by GIR systems. We motivate the need of a semantic-flavored query reformulation approach for geographic information retrieval systems and describe a GIR architecture where query reformulation focuses on i) grounding entities in the query, ii) selecting a reasoning strategy according to the user information need, and iii) generating a reformulated query containing answers and related entities for a more focused retrieval step. Reformulated queries obtain the answers by accessing a knowledge base.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Algorithms, Design, Evaluation

Keywords

Geographical Information Retrieval, Query Reformulation, Geographic ontology, Information management

1. INTRODUCTION

Geographic queries often have references to geographic and non-geographic entities Frequently such entities have a relationship link. For instance, the topic #1 from the GikiCLEF evaluation task [14], "List the Italian places where Ernest Hemingway visited during his life," expects a list of geographic places in Italy, which have a specific relationship with a given entity, in this case a person. This is an example of a real user information need with geographic restrictions, which requires a semantic query reformulation module with strong reasoning capabilities. It also requires a comprehensive knowledge base of geographic and non-geographic entities, to search for concrete answers of Italian places that match the conditions in the query (places visited by Hemingway). These answers can be used to reformulate the initial query for the final retrieval step, with optimised index weights giving a higher ranking scope to documents referring to one of such places.

Statistical query reformulation techniques, like blind relevance feedback (BRF), improve results in classic text retrieval systems [17]. In BRF, an initial query is reformulated by adding more terms, which are extracted from the top-n documents of an initial retrieval, and ranked using a weighting algorithm based on a term statistic approach. Some of these reformulated queries are in fact better descriptions of the initial information needs, yielding retrieval results containing more relevant documents. However, this method can also generate queries that drift from the initial query and add noise in the retrieval step [12].

Regardless of the unpredictable effect of BRF on the retrieval results, it is highly questionable if expanded terms like "rome", "milan", "bell" or "tolls" are a better description of the information need from the above geographic query. In our opinion, for GIR we need a different query reformulation approach, a semantic-flavored approach that will find correct answers like "Fossalta di Piave", "Torcello" or "Acciaroli" for that query, and will include them in a reformulated query for the document retrieval.

We present a GIR architecture with a semantic query reformulation approach that does not work directly with terms, but with the *entities* that are represented by those terms exploring all the information available on such entities, namely properties and relationships among them. Earlier GIR approaches had already done this for terms designating geographic places, spatial relationships or geographic feature types, but only with the purpose of determining the geographic scope of queries [3]. However, seeing a geographic query as composed of a main subject and geographic criteria that can be handled separately is not always a viable strategy [2]. The proposed query reformulation approach adapts its reasoning strategy according to the type of a geographic query, as we detail later on this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GIR'10, 18-19th Feb. 2010, Zurich, Switzerland
Copyright 2010 ACM ISBN 978-1-60558-826-1/10/02 ...$10.00

The paper is organised as follows: Section 2 details the proposed architecture for GIR. Section 3 dissects the semantic query reformulation approach. Section 4 describes the document annotation tool. Section 5 describes the knowledge base that provides access to several knowledge resources. Section 6 describes the index generation and how they are used by the document retrieval & ranking module. Section 7 illustrates query reformulation strategies for some types of geographic queries, and Section 8 concludes the paper with additional remarks.

2. SYSTEM OVERVIEW

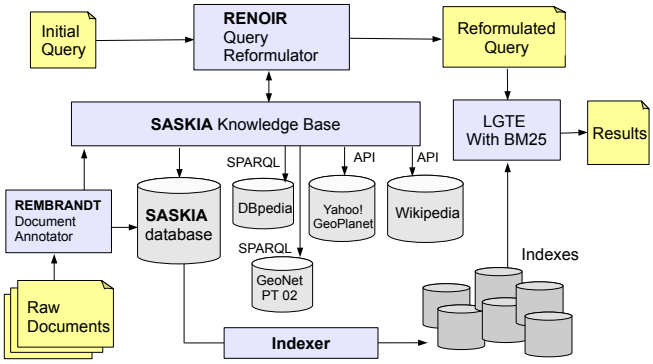

Figure 1: Overview of the GIR architecture

Figure 1 presents the GIR architecture for implementing semantic query reformulation of geographic queries. The prototype has five main modules: i) a semantic query reformulation module, Renoir, handles and reformulates user queries (detailed in Section 3), ii) a document annotator tool, Rembrandt, recognizes and grounds all entities from documents (detailed in Section 4), iii) a knowledge base, Saskia, is the access point for all knowledge resources (detailed in Section 5), iv) an indexer, which generates a standard term index and selective indexes for each entity type, and v) a retrieval and ranking module, the LGTE (Lucene with GeoTemporal Extensions) [10]. The indexer and LGTE are detailed in Section 6.

As knowledge resources, we use a local copy of the English and Portuguese Wikipedia snapshots (article texts and SQL dumps), a local copy of the DBpedia dataset [1] and the geographic ontologies GeoNetPT-02 for the Portuguese territory [9], both loaded in a OpenLink Virtuoso triplestore server, and the Yahoo!'s GeoPlanet™web-service [18].

3. QUERY REFORMULATION

Figure 2 details the structure of Renoir. The initial task of Renoir is performed by a *question interpreter*, which parses queries into object representations (*question objects*). From there, the *question reasoner* selects the best strategy to reformulate the query, which may include performing information extraction steps over the knowledge resources.

3.1 Question interpreter

The goal of the question interpreter (QI) is to recognise entities from the query and ground them using unique identifiers

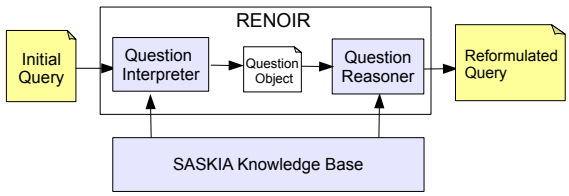

Figure 2: RENOIR query reformulation module

(such as DBpedia resource URLs or GeoPlanet™WOEIDs – Where On Earth IDs [7]), thus generating a machine-interpretable object (the question object) that encapsulates the entities that describe the user's query.

The question object, which is used to select the right reasoning strategy for each query, is composed of the following attributes:

Subject, list of expected answer types. The subject can be grounded as i) a DBpedia resource that has a property `rdf:type` for a value `skos:Concept` [11], ii) a DBpedia ontology class, or iii) a semantic classification as defined in the HAREM categorization [15], in this preferential order. (HAREM is an evaluation contest for named entity recognition in Portuguese – `http://www.linguateca.pt/HAREM/`.)

Expected Answer Type (EAT), properties that the set of answers must have, if the query is formulated as a question. For example, the EAT can be assigned to the DBpedia resource `http://dbpedia.org/page/Category:American_journalists`, the ontology class `http://dbpedia.org/ontology/Writer` or the HAREM classification `PERSON/INDIVIDUAL`.

Conditions, list of filtering criteria to the subject, such as a geographic scope or a temporal expression.

The question object is generated by applying a set of pattern rules over the query terms. Take for instance a query in the form of a question "*Which Romanian writers were born in Bucharest?*". The QI would perform as follows:

Ground the subject. A first set of pattern rules detect the terms that define the subject; in the given example, the terms "*Romanian writers*" would be captured and grounded to `http://dbpedia.org/resource/Category:Romanian_writers`, which is a DBpedia resource derived from the corresponding Wikipedia's category page, and has a property `rdf:type` with the value `skos:Concept`.

Ground the EAT. After the subject is grounded, pattern rules determine the EAT according to the question type and the subject. For queries with a "*Which X*" pattern as in the given example, the EAT is assigned to the subject, that is, the answers must have a `skos:subject` property with the subject's resource as the value, meaning that all the answers must be about Romanian writers. If the subject

cannot be mapped to a DBpedia resource, it is mapped to a DBpedia ontology class or a HAREM category, which may be assigned to the EAT. Suppose that the QI failed to ground "*Romanian writers*" to a DBpedia resource; using a simple term mapping hash, the term "*writers*" triggers a mapping of the EAT to the `http://dbpedia.org/ontology/Writer` class. Lastly, if the QI cannot map to a DBpedia ontology class, a generic category/type `PERSON/INDIVIDUAL` is used instead.

Ground conditions. In the given example, there is a condition that filters the correct answers from an initial list of Romanian writers to those who were born in the city of Bucharest. Pattern rules triggered by the terms "*were born in Bucharest*" generate a new condition with a property grounded to the `dbpedia-owl:birthPlace` property (`http://dbpedia.org/ontology/birthplace`, and a referent entity grounded to the DBpedia resource `http://dbpedia.org/resource/Bucharest`, which in this case acts also as a geographic scope for the query.

In a related work on semantic interpretation of queries, Hartrumpf and Leveling [6] employ a technique called query decomposition where they synthesize the initial query into a list of simpler sub-queries which are easier to answer. In the given example, the query could be decomposed to "*Name Romanian writers*" and "*Did X was born in Bucharest?*".

3.2 Question reasoner

The question reasoner (QR) processes the grounded entities within the question object, aiming to resolve them into a list of answers. Depending on the elements present in the question object, the QR task decides the best strategy to obtain those answers, which consists of a list of queries made to the SASKIA knowledge base. In the given example, for a question object with an EAT given by `Category:Romanian_writers`, a single condition described by a property `dbpedia-owl:birthPlace` and a referent geographic entity `http://dbpedia.org/resource/Bucharest`, the QR strategy solves the question by issuing the following SPARQL query to the DBpedia dataset:

```
SELECT ?RomanianWriters WHERE {
 ?RomanianWriters
    skos:subject
 <http://dbpedia.org/resource/Category:Romanian_writers> .
 ?RomanianWriters
    <http://dbpedia.org/ontology/Person/birthPlace>
 <http://dbpedia.org/resource/Bucharest>
}
```

In the above example, if the condition was "*born in Romania*", the QR would rewrite the condition as "*born in ?X*" and "*?X in Romania*", and would use a geographic ontology to provide the necessary geographic knowledge to resolve the geographic scope of a query, as in:

```
SELECT ?RomanianWriters WHERE {
 ?RomanianWriters
    skos:subject
 <http://dbpedia.org/resource/Category:Romanian_writers> .
 ?RomanianWriters
    <http://dbpedia.org/ontology/Person/birthPlace>
 ?Romanian_place .
 ?Romanian_place
    sameAs
 ?ontology_Romanian_place .
 ?ontology_Romanian_place
    ontology:partOf
 ?ontology_Romania .
 ?ontology_Romania
    sameAs
 <http://dbpedia.org/resource/Romania>
}
```

4. DOCUMENT ANNOTATOR

REMBRANDT (`http://xldb.di.fc.ul.pt/Rembrandt/`) is a named-entity recognition (NER) software which is used to annotate documents by classifying named entities (NEs) and grounding them into identifiers composed by Wikipedia and DBpedia URLs.

REMBRANDT classifies NEs using the 9 main categories and 47 sub-categories defined by the second edition of HAREM [15]: `PERSON`, `ORGANIZATION`, `PLACE`, `DATETIME`, `VALUE`, `ABSTRACTION`, `EVENT`, `THING` and `MASTERPIECE`. Its classification strategy relies on mapping NEs to their corresponding Wikipedia and DBpedia pages, using DBpedia's ontology classes and Wikipedia categories to infer the semantic classification. Afterwards, REMBRANDT applies a set of manually generated language-dependent rules, which represent the internal and external evidence for NEs for a given language, as in "*city of X*" or "*X, Inc.*" The set of rules disambiguates NEs with more than one semantic classification, and also classifies NEs that were not mapped to a Wikipedia or DBpedia page.

4.1 Generating geographic and temporal signatures

REMBRANDT generates *geographic signatures* and *temporal signatures* for tagged documents. We designate a geographic signature of a document as the surrogate of the NEs found in the document that were grounded as a geographic place, where each place is expanded upwards to the country level, following the strategy proposed by Li et al. [8].

Figure 3 illustrates a geographic signature of a document. We have experimented using geographic signatures document and query surrogates before [4], but we just used geographic term with no grounding information. A geographic signature is composed by a list of `<Place>` elements, which has a `count` attribute that stores the document frequency of that place, and a WOEID (or also a GeoNetPT-02 ID if the place is within the Portuguese territory) as an identifier. Each `<Place>` element contains the different NEs used in the document to designate the place (`<NE>`), the place's entity name (`<Entity>`), HAREM's semantic classifications (`<Type>` and `<Subtype>`), the DBpedia class (`<DBpediaClass>`) and the ancestor's WOEIDs and entity name (`<Ancestor>`).

Figure 4 illustrates a temporal signature of a document generated by REMBRANDT. A temporal signature is the surrogate of the NEs of category `DATETIME` found in the

```
<GeoSignature totalcount="4">
  <Doc id="571048" lang="en">
  <Place count="2" woeid="1467052">
     <NE>Harare</NE>
     <Entity>Harare</Entity>
     <Type>@HUMAN</Type>
     <Subtype>@DIVISION</Subtype>
     <DBpediaClass>Area</DBpediaClass>
     <Ancestor woeid="56048529">Harare</Ancestor>
     <Ancestor woeid="24550731">Harare</Ancestor>
     <Ancestor woeid="23425004">Zimbabwe</Ancestor>
  </Place>
  <Place count="2" woeid="23425004">
     <NE>Zimbabwe</NE>
     <Entity>Zimbabwe</Entity>
     <Type>@HUMAN</Type>
     <Subtype>@COUNTRY</Subtype>
     <DBpediaClass>Country</DBpediaClass>
  </Place>
</GeoSignature>
```

Figure 3: Geographic signature of a document.

```
<TimeSignature>
 <Doc id="523634" lang="en">
<Time count="1">
   <NE id="3645" lang="en">2006</NE>
   <TimeGrounding>!:Y+2006</TimeGrounding>
   <Index type="date">2006</Index>
 </Time>
 <Time count="1">
   <NE id="3646" lang="en">27th May, 2006</NE>
   <TimeGrounding>!:Y+2006M05S27</TimeGrounding>
   <Index type="date">20060527</Index>
 </Time>
 <Time count="1">
   <NE id="3648" lang="en">28th May, at 15h30</NE>
   <TimeGrounding>!:M05D28h15m30</TimeGrounding>
   <Index type="date">20060528</Index>
 </Time>
</TimeSignature>
```

Figure 4: Temporal signature of a document.

document that can be grounded into a temporal expression. Each distinct NE is represented in a `<Time>` element, with a `count` attribute storing its document frequency. The `<NE>` element contains the NE terms. The `<TimeGrounding>` element contains the temporal expression that grounds the NE in an easier format to process. The `<Index>` element is generated by rounding temporal expressions to a day threshold, and by filling the missing elements by looking at the surrounding time expressions and selecting the most popular values. For example, in the time expression "28th May, at 15h30" from the time signature of Figure 4, the year 2006 is added, inferred by the other time expressions in the document, which explicitly refer to that year.

5. KNOWLEDGE BASE

Figure 5 describes the SASKIA knowledge base. SASKIA is the access point to the information extracted from the document collection, which is stored in a database with the grounded information about all entities found. SASKIA's API also facilitates access to external knowledge resources, such as Wikipedia and DBpedia.

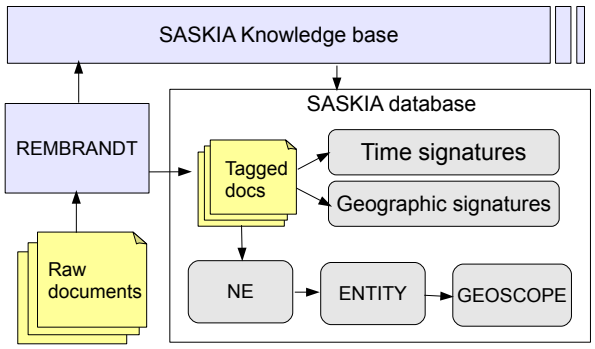

Figure 5: The Saskia knowledge base

For each document tagged by REMBRANDT, each NE is stored in a NE table, together with its terms and its semantic classification. If the NE was successfully grounded, that information is stored in an ENTITY table. This NE/ENTITY separation is needed, as a given entity can be referred to by several NEs (for example, the entity "*United States of America (country)*" can be represented by the NEs "US", "USA", "United States"). Likewise, the NE table distinguishes between NEs with the same terms and/or same semantic classification, but with different grounded entities. For example the NE "London" refers to different cities in Canada and in the UK (the disambiguation was previously made by REMBRANDT). Moreover, RENOIR can therefore have access to a list of different NE representations for a given entity, which can be useful in the query reformulation step, to circumvent the different designations given by different authors in the document collection.

SASKIA's database has also a GEOSCOPE table, which encompasses all grounded information for entities that are geographic places, such as centroids, bounding boxes, ancestors or neighbors, as well as WOEIDs and GeoNetPT-02 identifiers, and is used for the generation of geographic signatures.

6. INDEX GENERATION AND DOCUMENT RETRIEVAL

The retrieval and ranking step is performed by LGTE [10], a Lucene package with extensions for geographic and temporal ranking measures, using a BM25 weighting scheme implementation [13]. LGTE has three kinds of indexes: i) an index of document terms, ii) an index for NEs, and iii) an index of document signatures.

The index for document terms is a standard inverted index used in classic text retrieval, providing a baseline for all retrievals if RENOIR fails to perform a query reformulation. The index for NEs is an inverted index for the NE terms within a given semantic classification. For instance, an index for NEs of category `PERSON` includes only the terms included in NEs that were classified as `PERSON`, along with their grounded entity representation. If a document contains the NE "Hemingway" grounded to the entity "Ernest Hemingway", then both designations are indexed.

Lastly, the index for document signatures is the base for LGTE's geographic and temporal similarity rankings. The

geographic signature index combines the entity's WOEID identifier with its term, to distinguish between different places with identical placenames, and includes the ancestors information. Taking the geographic signature from Figure 3, its index entries have the following format:

```
GeoscopeID                : DocID  <DocFreq>
Harare@WOEID-1467052      : 571048 <2>
Harare@WOEID-56048529     : 571048 <2>
Harare@WOEID-24550731     : 571048 <2>
Zimbabwe@WOEID-23425004 : 571048 <4>
```

As "Harare" has a document frequency of 2, its ancestors (`Harare@WOEID-1467052` as the county, `Harare@WOEID-56048529` as the province) also have that document frequency value, while `Zimbabwe@WOEID-23425004` as the country adds other two occurrences, totaling a document frequency value of 4.

With this index format, LGTE can generate a simple geographic similarity measure for queries with a geographic scope. For example, if a query is grounded to Zimbabwe as a country, then all documents that refer to cities in Zimbabwe are eligible to be retrieved, since they all have a reference to `Zimbabwe@WOEID-23425004` in the geographic signature index, added as an ancestor.

The temporal signature index presents a similar format. For the example in Figure 4, the index entries are:

```
Time     : DocID  <DocFreq>
2006     : 523634 <1>
20060527 : 523634 <1>
20060528 : 523634 <1>
```

To compute temporal similarity measures, LGTE can use wildcards during index search. For instance, a query with temporal restriction of "May 2006"is grounded as `200605*`, which encompasses all documents with references to a date within that month, including `20060527` and `20060528`.

As LGTE can combine the scores given by multiple indexes for a single retrieval, we can tune the index weights for each reformulated query. For instance, a query containing a reference to the entity "*Ernest Hemingway (person)*" could trigger a boost on the weight value of the `PERSON` NEs index, so that documents referring to Hemingway as a person will have a higher ranking score from those with a reference to Hemingway in a non-person role.

7. QUERY REFORMULATION STRATEGIES

Santos and Chaves presented a classification for geographic topics which summarises the type of geographic queries that GIR systems should handle [16]. This classification was later used by GeoCLEF organizers as topic selection guidelines [5].

Renoir has optimised strategies for the first five query classifications, as the other classifications "*geographical relations among places*" (e.g., *how are the Himalayas related to Nepal?*), "*geographical relations among (places associated to) events*" (e.g., *Did Waterloo occur more north than the battle of X?*) and "*relations between events which require their precise localization*" (e.g., *was it the same river that flooded last year and in which killings occurred in the XVth century?*) are out of the scope of the research work introduced in this paper.

We illustrate below the Renoir strategies using the GeoCLEF examples [5]:

1. non-geographic subject restricted to a place. (e.g., *music festivals in Germany*). Renoir grounds the EAT of the query as in the DBpedia ontology `http://dbpedia.org/ontology/MusicFestival`, and the query scope is grounded to the entity `Germany@WOEID-23424829`. Renoir searches the knowledge base for instances of music festivals that occur in Germany, using the following SPARQL query:

```
SELECT ?MusicFestival WHERE {
 ?MusicFestival
    rdf:type
 <http://dbpedia.org/ontology/MusicFestival> .
 ?MusicFestival
    <http://dbpedia.org/ontology/location>
 <http://dbpedia.org/resource/Germany>
}
```

All answers are added to the reformulated query. Since the majority of the answers are events, the LGTE will boost the index weight to `EVENT` NEs, and will also use the geographic signature index to give additional scores for documents with Germany within their scope. The reformulated query could be similar to:

term: music festivals germany "wacken open air" "zappanale" "summerjan" "summer breeze open air"
event: "Wacken Open Air" "Zappanale" "Summerjan" "Summer Breeze Open Air"
place: Germany
geoscope: `Germany@WOEID-23424829`

where each term is assigned to a target index. Entities with multiple terms are wrapped with double quotes, so that LGTE matches documents that have those term in the same order. For instance, the terms "wacken open air" will be matched for the term and `EVENT` NEs indexes, the term "germany" will be matched to the term and `PLACE` NEs indexes, while the term Germany@WOEID-23424829 will be matched only to the geographic signature index.

2. geographic subject restricted to a place. (e.g., *cities in Germany*). Geographic queries of this type are similar to the previous case, although a geographic ontology would be the preferable knowledge resource to use. If the query scope was within the Portuguese territory, for instance, *cities in Portugal*, Renoir would issue the following SPARQL code to GeoNetPT-02:

```
SELECT ?city WHERE {
  ?city
     gn:type
  gnpt02:localidade-ATLOC .
  ?country
     gn:type
  gnpt02:pais-ATPAI .
  ?country
     dc:title
  "Portugal" .
  ?country
     gnpt:hasPart
  ?city .
}
```

Outside the Portuguese territory, RENOIR would use DBpedia, since the GeoPlanet™web service does not have a SPARQL query endpoint. The reformulated query could be similar to:

term: cities germany berlin dusseldorf munich leipzig hamburg
place: Berlin Dusseldorf Munich Leipzig Hamburg
geoscope: `Germany@WOEID-23424829` `Berlin@WOEID-638242` `Dusseldorf@WOEID-646099` `Munich@WOEID-676757` `Leipzig@WOEID-671072` `Hamburg@WOEID-656958`

3. non-geographic subject associated to a place. (e.g., *independence of Quebec*). This type of queries can be included in the set of queries where RENOIR cannot ground the EAT, grounding only the query scope to the Québec region. In these cases, the retrieval step will use the term and the geographic signature indexes for the reformulated query.

4. geographic subject with non-geographic restriction. (e.g., *rivers with vineyards*). This type of geographic queries have a geographic place as EAT, but with no explicit geoscope. In this case, RENOIR grounds the EAT as a geographic feature (given by DBpedia's ontology property `http://dbpedia.org/ontology/River`).If there is referent entity (as in *rivers that Ernest Hemingway visited*), a reasoning round could be performed to obtain a list of answers for the reformulated query. If not, RENOIR performs a document retrieval with no additional terms in the reformulated query.

5. non-geographic subject that is a complex function of place. (e.g., *winners of Eurovision Song Contest 2009*). In these types of geographic query, there is an implicit geographic scope and/or a temporal scope that is associated to an entity, as for instance an event or an organization. In this example, RENOIR would ground the single entity to the DBpedia resource `http://dbpedia.org/page/Eurovision_Song_Contest_2009`, and classify it as `EVENT`. To resolve the implicit geographic end temporal scope, the following DBpedia query can be issued:

```
SELECT ?geo, time, WHERE {
  <http://dbpedia.org/page/Eurovision_Song_Contest_2009>
     <dbpedia.org/ontology/location>
  ?geo .
  <http://dbpedia.org/page/Eurovision_Song_Contest_2009>
     <dbpedia.org/ontology/date>
  ?time
}
```

the reformulated query could be similar to:

term: winners eurovision song contest 2009
event:"Eurovision Song Contest 2009"
geoscope: `Moscow@WOEID-2122265`
timescope: 2009*

8. FINAL REMARKS

We presented a semantic query reformulation approach which we believe is more suited for geographic queries than query reformulation approaches based on term statistics such as blind relevance feedback. We pick for each query a reasoning strategy according to the type of geographic query and the entities grounded in it, and we generate a reformulated query that may include a list of answers that match the user's information need, a grounded query scope or a temporal scope. In the retrieval step, the retrieval & ranking module selects and adjust weights indexes so that the document ranking scores include a geographic and temporal similarity measure.

We are developing a GIR prototype that has all the modules fully developed except RENOIR, in which the grounding algorithms are only capable of parsing simple geographic queries and the reasoning algorithms are still limited to basic SPARQL queries. When the prototype is fully developed, we will compare its performance against an IR system with blind relevance feedback, and we will measure the performance gains using the GeoCLEF evaluation resources (topics, collections and judgements).

The performance of the proposed semantic query reformulation is highly dependent on the information coverage provided by the knowledge resources. One problem is to pick the wrong set of SPARQL queries or API calls that do not get the right answers for a given query, another different problem is if the set of SPARQL queries or API calls are right, but there are no answers because they are not yet covered in the DBpedia dataset or in the geographic ontologies. Nonetheless, while the performance gain of the proposed GIR approach can be moderate at start for some queries, due to a limited coverage of the knowledge resources, these resources will evolve over time and their detail and coverage of known entities will increase.

9. ACKNOWLEDGMENTS

This work is supported by FCT for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006) and a PhD scholarship grant SFRH/BD/45480/2008.

10. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, number 4825 in LNCS, pages 722–735. Springer, 2007.

[2] N. Cardoso and D. Santos. To Separate or Not to Separate? Reflections about Current GIR Practice. In *Workshop on Novel Methodologies for Evaluation in Information Retrieval, NMEIR 2008 (held at ECIR 2008)*, Glasgow, Scotland, 30th March 2008.

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

[4] N. Cardoso, M. J. Silva, and D. Santos. Handling Implicit Geographic Evidence for Geographic IR. In J. G. Shanahan, S. Amer-Yahia, I. Manolescu, Y. Zhang, D. A. Evans, A. Kolcz, K.-S. Choi, and A. Chowdhury, editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 1383–1384, Napa Valley, CA, USA, October 26-30 2008. ACM.

[5] F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, G. D. Nunzio, and N. Ferro. Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF. In *Proceedings of the 1st International Workshop on Evaluating Information Access, EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)*, Tokyo, Japan, May 15 2007.

[6] S. Hartrumpf and J. Leveling. GIRSA-WP at GikiCLEF: Integration of Structured Information and Decomposition of Questions. In *Cross Language Evaluation Forum: Working Notes for CLEF 2009*, Corfu, Greece, 30 September - 2 October 2009.

[7] Y. Inc. Yahoo! GeoPlanet Key Concepts. `http://developer.yahoo.com/geo/geoplanet/guide/concepts.html`, 2010. accessed on January 2010.

[8] Y. Li, A. Moffat, N. Stokes, and L. Cavedon. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In R. Purves and C. Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*, pages 17–22. Department of Geography, University of Zurich, 2006.

[9] F. J. Lopez-Pellicer, M. Chaves, C. Rodrigues, and M. J. Silva. Geographic Ontologies Production in Grease-II. Technical Report DI/FCUL TR 09-18, Faculty of Sciences, University of Lisbon, November 2009. DOI 10455/3256.

[10] J. Machado. LGTE: Lucene Extensions for Geo-Temporal Information Retrieval. In *Workshop on Geographic Information on the Internet Workshop (GIIW)*, Toulouse, França, 2009.

[11] A. Miles, B. Matthews, M. Wilson, and D. Brickley. SKOS Core: Simple Knowledge Organisation for the Web. In *DCMI '05: Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications*, pages 1–9. Dublin Core Metadata Initiative, 2005.

[12] 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'98)*, pages 206–214, Melbourne, Australia, August 24-28 1998. ACM Press.

[13] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 21–30, Gaithersburg, MD, USA, 1992.

[14] D. Santos and L. M. Cabral. GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia. In F. Borri, A. Nardi, and C. Peters, editors, *Cross Language Evaluation Forum: Working notes for CLEF 2009*, 30 September - 2 October 2009.

[15] D. Santos, P. Carvalho, H. Oliveira, and C. Freitas. Second HAREM: new challenges and old wisdom. In A. Teixeira, V. L. S. de Lima, L. C. de Oliveira, and P. Quaresma, editors, *Computational Processing of Portuguese Language, 8th International Conference (PROPOR'2008), September 8-10, Aveiro, Portugal. Proceedings*, number 5190 in LNCS, pages 212–215. Springer, 2008.

[16] D. Santos and M. Chaves. The Place of Place in Geographical IR. In R. Purves and C. Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*. Department of Geography, University of Zurich, 2006.

[17] J. Xu and W. B. Croft. Query Expansion Using Local and Global Document Analysis. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 4–11, Zurich, Switzerland, August 18-22 1996.

[18] Yahoo! Geoplanet™. http://developer.yahoo.com/geo/geoplanet/.