Every Document Has A Geographical Scope

G. Andogah[a,b], G. Bouma[b], J. Nerbonne[b]

[a]*Department of Computer Science, Gulu University, P. O. Box 166, Gulu, Uganda*
[b]*Alfa Informatica, University of Groningen, 9700 AB Groningen, The Netherlands*

Abstract

It is a useful premise to assume that every document in a collection and every query issued to an information retrieval (IR) system is geography-dependent. If one can determine what area an article is about (i.e., its' geographical scope), this information can be used to improve the accuracy with which people, places and organizations named in the article can be located. More importantly, geographical scopes of documents may be exploited to improve the performance of IR systems against geography-dependent user queries by tuning relevance ranking and query expansion strategies with scope metadata. We want to answer the following pertinent questions to ascertain the usefulness of geographical information in improving retrieval accuracy: (1) how far can geographical information in queries and documents improve retrieval accuracy of IR systems when answering geography-dependent queries; and, (2) how effectively can geographical information in queries and documents be utilized to improve the quality of relevance ranking in geographical IR domain. This paper outlines strategies to determine the geographical scope of documents, and describes methods to utilize scope information to improve the performance of toponym resolution, relevance ranking and query expansion.

Keywords: Geographic information retrieval, geographical scope resolution, toponym resolution, query expansion, relevance ranking

1. Introduction

It is a useful premise to assume that every document in a collection and every query issued to an information retrieval (IR) system is geography-dependent. If we can globally determine what area an article or a document is about (i.e., its' geographical scope), we can reasonably assume that people, places and organizations named in the article are located in the area. Formally, we define a *geographical scope of a document* as the region or area for which the document is geographically relevant. The scope[1] metadata of documents can be exploited to

Email addresses: g.andogah@gu.ac.ug (G. Andogah), g.bouma@rug.nl (G. Bouma), j.nerbonne@rug.nl (J. Nerbonne)

[1] The usage of the term *scope* in this paper is synonymous to *geographical scope.*

(1) categorize documents by scope, (2) improve performance of toponym resolution, and, (3) tune relevance ranking and query expansion strategies to improve retrieval performance of IR systems against geography-dependent queries.

This paper outlines strategies to determine the geographical scope of documents, and describes methods to utilize scope information to improve performance of toponym resolution, relevance ranking and query expansion. The basic argument of this paper is simple. If one first detects the geographical scope of an entire document, then several other processes in geographical information retrieval may be improved.

This paper assumes that named entity recognition (NER) software is available in order to focus on toponym resolution, i.e., identifying the places that place names refer to. This focus is justified since state-of-the-art NER (place and person name tagging) already performs quite well on English text [1]. We therefore opted to use evaluation datasets where the names of places and persons [1, 2] had already been identified, which allowed us to evaluate scope and toponym resolution procedures independently (of NER accuracy). To evaluate the proposed relevance feedback and ranking procedures, an off-the-shelf NER tool was used to annotate the dataset.

Summary of Contribution. A brief summary of the contribution of the paper is as follows:

1. We defined and modelled geographical scopes as a kind of document (see Section 3.1 and 3.2).
2. We resolve geographical scope of a document prior to disambiguation of places mentioned inside the document.
    - The toponym-based scope resolution scheme achieved performance of 79% on human annotated news articles (see Section 3.3.1).
    - To the best of our knowledge, this work is the first to attempt to use the knowledge of persons mentioned in documents to infer geographical scope of document. The scheme achieved performance of 58% on human annotated news articles (see Section 3.3.2).
3. We base the toponym resolution procedure grounded on geographical scopes assigned to documents. The procedure achieved performance of 71% - 80% on human annotated news articles (see Section 4.1 and 4.2).
4. We show that geographic metadata (e.g., toponyms, scopes) can improve performance of IR systems when answering geography motivated queries (sec Section 5 and 6).
    - We introduce a relevance feedback procedure that selects toponyms within the scopes of relevant documents to expand geography motivated queries achieved an improvement of 9.5% over standard IR system (see Section 5.3).
    - We introduce a relevance ranking metric that exploits a ranked list of geographical scopes assigned to queries and documents achieved an improvement of 8.9% against standard IR system (see Section 6.4).

	Adjacent-to	Part-of	Geo-distribution	Feature importance	Toponym frequency	Resolved toponyms
[4]				*		*
[5]	*	*				*
[6]		*			*	*
[7]		*	*		*	*
[8]			*		*	*
[9]		*	*		*	*
Mahali	*	*	*	*	*	

Table 1: Scope resolution techniques in literature.

2. Related Work

2.1. Scope Resolution

A document's geographical scope(s) is/are the geographical regions or areas which the document is about. The geographical scopes of a document can help in retrieving documents by imposing geographical restrictions on the search query [3], and can also be used in the toponym resolution task [4, 2]. Our scope resolution system assigns scopes at six levels, namely, continent, continental regional grouping, country, country compass regions (e.g., south-east Spain), province[2] and province compass regions. This allows us to introduce more fine-grained scope metadata to drive toponym resolution, relevance ranking and query expansion routines than previously attempted. This subsection reviews recent works in geographical scope resolution.

Table 1 shows geographical information explored by different authors in the literature to determine the geographical scope of a document. The abbreviations used as column titles stand for:

- *adjacent-to* relationship among geographical scopes or concepts in a geographical knowledge base. For example, *Belgium* ↔ *Netherlands* ↔ *Germany.* The *adjacent-to* relationships are considered less important than the *part-of* relationships.
- *part-of* relationship among geographical scopes or concepts in a geographical knowledge base. For example, *The Hague* ↦ *South Holland* ↦ *Nether-*

[2] Province here refer to first administrative division of a country, e.g., Groningen Province, New York State, etc.

lands ↦ *Europe* relationship. The *part-of* relationships are considered more important than the *adjacent-to* relationships.

- *geographical distribution* of places taking part in the definition of geographical scopes. This requires that (1) a significant fraction of all locations mentioned in the document are either the scope itself or locations within the scope, and, (2) the location references in the document are distributed smoothly across the scope. For example, a document with scope *New York State* is expected to mention *New York State* or locations within *New York State* more frequently than places belonging to other states or countries.

- *importance of geographical feature* determined by feature type and/or population size. The scope of a document is set to the country or region containing the most important unambiguous place names (e.g., capital cities and other major cities) identified in the document. For example, spotting *Rotterdam*, *The Hague* and *Amsterdam* in a document sets the scope of the document to the *Netherlands*.

- *toponym frequency* of occurrence in the document. Here the most commonly occurring place in the document dictate the scope of the document. For example, if *New York State* is mentioned more frequently than *Groningen Province*, the scope of the document is most likely the *New York State*.

- *resolved toponyms* are used to compute a document's geographical scope. Before an attempt is made to determine the scope of the document, the toponyms identified in the document are resolved to the location they refer to.

As show in Table 1, our system called *Mahali* [3] (see the bottom row in Table 1) combines all the geographical information used in literature reviewed with exception that we use toponyms before they are resolved to the locations they refer to. Scope resolution using unresolved toponyms is a unique feature of our approach where the previous approaches use resolved toponyms to determine scopes of documents [10].

2.2. Toponym Resolution

Many places on the surface of the earth share names – for example, *London (England)* and *London (Ontario)*, and many places also have multiple names; for example, the names the *Netherlands* and *Holland* refer to the same place. Toponym resolution is a process of assigning a toponym (place name) identified in text to a single non-ambiguous place on the earth's surface.

The following sub-section briefly reviews the state-of-the-art approaches in toponym resolution relevant to the work reported in this paper.

[3] All the components developed in the course of this work form part of our system called *Mahali*. Mahali means 'place' in Kiswahili.

2.2.1. Default Sense Heuristics

We define a default sense as the most likely sense in a given context, given that all the other parameters are constant for all the competing candidate senses. The likelihood of a candidate location being referred to is determined by the importance attached to it, and the following parameters have been used as indicators of importance:

- *land surface area* – selects the referent with the largest land surface area as the place referred to in the text [11, 12, 4].
- *hierarchy distance* – selects the place highest in the hierarchy of regions as the place referred to in the text [13, 14]. For example, *Holland (Europe)* will normally be preferred over *Holland (Michigan).*
- *place type* – selects a place in the order of place type importance: country → capital → city → town → village [13, 14, 15].
- *corpus popularity* – selects a place that occurs more commonly in the document collection as the place being referred to [16, 14]. For example, *Boston (US)* is preferred over *Boston (UK).*
- *population* – selects a place with the largest population as the place being referred to [11, 6, 4, 14].

2.2.2. Pattern Matching and Hierarchy Overlap

This approach exploits local pattern matching, the hierarchical part-of relation and spatial distance. In the literature the following techniques have been used:

- *feature type qualifier* – scans for the feature type of the target toponym in the text, e.g., *province of* Groningen, *capital city of* Kampala, Kilimanjaro *Mountain*, etc. The candidate referent with the matching type is selected [14, 15].
- *text and hierarchy overlap* – computes overlap between toponyms in the text and spatial hierarchy relations [13, 14]. For instance, a text containing toponyms *London, Southern Ontario, Canada* grounds toponym *London* → *London (Ontario).*
- *country scope restriction* – assigns a country scope to documents, and all the ambiguous toponyms are treated as belonging to the country assigned to the document [4].
- *smallest polygon* – resolves toponyms recognized in text to the smallest polygon that completely grounds the whole set [6, 17, 7]. Any other ambiguity is resolved using local pattern matching. This in a way is a scope restriction technique.
- *spatial distance* – decision is made on the basis of how close a candidate referent is to the non-ambiguous referents. The referent closest to all the non-ambiguous referents is chosen [11, 7, 4].

	Land surface area	Hierarchy distance	Place type	Corpus popularity	Population	Type qualifier	Text-hierarchy overlap	Country scope	Smallest polygon	Spatial distance	One Referent
[11]	*				*					*	
[4, 12]	*							*		*	
[13]		*	*				*				
[14]		*	*	*	*	*	*				
[16]				*							
[6]					*				*		
[15]						*					
[17]									*		*
[7]									*	*	
Mahali			*		*			*			*

Table 2: Toponym resolution techniques in literature.

2.2.3. One Referent per Discourse

The *one referent per discourse* approach assumes one and only one meaning for a toponym in a discourse [17]. Any subsequent mention of the same toponym is assumed to convey the same meaning as the previous meaning.

The toponym resolution component of the Mahali system described in this paper borrows the following existing techniques and ideas (see *Mahali* row in Table 2): place type hierarchy, population size, scope restriction concept, default sense, and one referent per discourse. The unique feature of our approach is that all the procedures are performed within the framework of the scope of a document using an elaborate list of scopes assigned by our scope resolution component (see Section 3).

2.3. Query Expansion

Query expansion adds words or phrases deemed synonymous or closely related to user query terms with the view of retrieving more relevant documents. Query expansion techniques are categorized into global methods and local methods [18, 19]. Global methods expand or reformulate query terms independent of the query and results returned from the query. Global methods include: (1) query expansion with a thesaurus, (2) query expansion via automatic thesaurus generation, and (3) techniques like spell checking. On the other hand, the local methods expand query terms relative to the top ranked documents returned as a response to the initial or previous query.

2.3.1. Knowledge-based Expansion

Two types of knowledge-based geographical query expansion schemes exist, the *term-based* and *footprint-based* expansion. The term-based geographical query expansion involves adding geographical terms from a geographical knowledge source to the original query with the view of retrieving additional relevant documents within the user's geographical area of interest [20, 21, 22, 23]. The footprint-based query expansion uses a geographical coordinate system to perform query expansion. All identified place names in the query are translated to their corresponding geographical coordinates such as the latitude/longitude coordinates [24].

2.3.2. Relevance Feedback Expansion

Relevance feedback is a popular technique to perform query expansion. Query expansion terms are obtained from documents returned as responses to the previous search query. Three types of relevance feedback exist: *blind or pseudo-feedback*, *explicit feedback* and *implicit feedback*. Explicit and implicit feedback have not been used to perform geographical query expansion in the literature. In this paper we explore blind and explicit feedback for geographical query expansion task.

The blind or pseudo-relevance feedback method assumes that the top n documents returned as a response to the query are relevant [19]. Usually the top m most frequent terms from the top n documents are added to the previous search query, and then resubmitted to perform new search over the collection. In [22] standard blind relevance feedback improved performance for a German geographical IR task. The most improved queries added mostly proper names and word variations and very few irrelevant words.

In explicit feedback a user classifies documents returned as relevant or non-relevant [19]. The terms from the relevant documents are used to expand the previous query. A similar approach can be applied to perform geographical term expansion where the user is asked to classify returned geographical scopes as relevant or non-relevant. Next, places found within relevant geographical scopes are used to expand the geographic terms of the query.

Implicit relevance feedback automatically learns from the searcher by observing her/his preferences during searching. Based on the learned model it performs relevance feedback expansion on the user's query to retrieve more relevant documents [25, 26, 19]. A number of user behaviours have been used as sources of implicit feedback: reading time, saving, printing and selecting [27].

2.4. Relevance Ranking

The objective of relevance ranking in geographical IR is to present the user with a ranked list of documents satisfying both the non-geographical and geographical criteria in the query. Recently a number of approaches have been proposed in the literature, and this subsection reviews some of the works.

2.4.1. Euclidean Distance

The *Euclidean distance* ranks documents by proximity between the query and the document geographical footprints. The shorter the distance between the query and document footprint, the more relevant a document is to the query location [28, 29, 30].

2.4.2. Extent of Overlap

The *extent of overlap* between the query and document footprint can be used to rank documents by geographic criteria. The greater the extent of overlap between the query and document footprint, the higher the relevance of the document to the query [31, 32, 33, 29].

2.4.3. Containment Relations

Two cases are defined as a *containment relation* [33, 31, 29] when: (1) the document footprint is inside the query footprint, or (2) the document footprint contains the query footprint. For case (1), the geographical score is assigned on basis of the ratio of document area to query area. On the other hand, the geographical score for case (2) is assigned based on the ratio of query area to document area. Geographical scores that approach zero indicate that the document is less relevant to the query's geographical criteria.

2.4.4. Query Footprint as Filter

All documents whose geographical footprint overlap with the query footprint are considered relevant. These documents are finally ranked according to their non-geographic scores [34, 33].

2.4.5. Geographical Scope Indexing

Scope indexing associates each scope (i.e., region of interest) to a list of documents concerning it, and in the opposite direction associates each document with a corresponding scope. Likewise, a query is associated with a list of scopes. Every document belonging to a scope is assigned a score based on how geographically relevant the document is to the scope. The assigned scores are manipulated by a ranking function to present a ranked list of geography-restricted relevant documents to the user [35].

2.4.6. Other Criteria

Apart from the above mentioned approaches to determine the geographical relevance of a document, the following criteria are also used: (1) travel time between query and document footprints, (2) boundary connectivity between query and document footprints, (3) number of intervening places between query and document footprints, and (4) place name emphasis in the document [29, 11].

3. Geographical Scope Resolution

Every document has a geographical scope either explicitly expressed or implied somewhere in the document. The resolution of geographical scope of a document is non-trivial as the scopes of documents are highly ambiguous. The process of automatically assigning geographical scopes to documents is called *geographical scope resolution.* Documents generally carry geographical clues to facilitate scope resolution process, among which we have – toponyms, adjectives of places and people, names of people, names of organization, language of the document, etc. This section reports on two strategies that exploits toponyms and anthroponyms to determine the scope of a document. The novel contribution of this paper is to show that a global assessment of a document's geographical scope improves the accuracy with which the document's components (such as location entities) may be understood geographically. The novelty of our toponym-based scope resolution strategy is that it uses unresolved toponyms as opposed to methods reported in the literature which use resolved toponyms to detect the scopes of documents. To the best of our knowledge, this paper is the first to attempt to use names of people to detect the geographical scopes of documents.

3.1. Using Toponyms

The toponym-based scope resolution strategy is grounded on Assumption 1.

Assumption 1. *Places of the same type or under the same administrative jurisdiction or adjacent-to each other are more likely to be mentioned in a given discourse unit. For example, a discourse mentioning 'The Netherlands' is more likely to mention places of the type country (e.g., United Kingdom, Uganda) or places under the jurisdiction of 'The Netherlands' (e.g., Amsterdam, Rotterdam) or places adjacent to 'The Netherlands' (e.g., Belgium, Germany).*

Assumption 1 is modelled as shown Fig. 1. The hollow diamond head arrows indicate *part-of* relations, and line head arrows indicate the *neighbour* relationships. The *target-region* (also called the *target-scope*) is the geographical scope or area being described by the model. The *target-region* can have one and only one *parent-region.* A *neighbour* shares the same *parent-region*, common border and place type with the *target-region.* The *child-region* (also called the primary administration division) and *primary-cities* are direct descendants of the *target-region.* The *child-child-regions* and *secondary cities* are the secondary administrative divisions and cities in the primary administration division. The *smallest-cities* are cities found both in the primary and secondary administrative divisions. [4] The *parent-region* shares a one-to-many relationship with the

[4] The terms *primary administrative division* and *secondary administrative division* are used in the context of the target region or scope. For example, the Dutch *Province of Groningen* is a primary administrative division in the scope of the *Netherlands*, but a secondary administrative division in the scope of *Europe*.

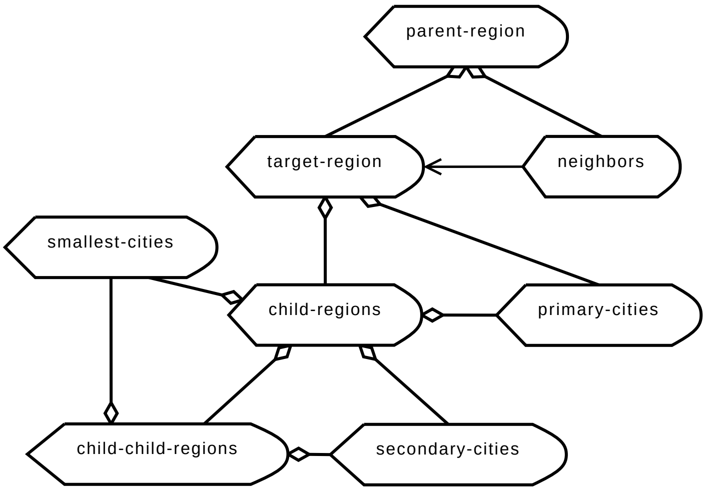

Figure 1: Data model for geographical scope modelling.

target-region and the *target-region*'s neighbour; the *target-region* shares a one-to-many relationship with its immediate *neighbour*, *child-region* and *primary-city*; *child-region* is in one-to-many relationships with the *child-child-regions*, *secondary-cities* and *smallest-cities*; and *child-child-region* shares one-to-many relationship with *primary-cities* and *smallest-cities.*

The contribution of each geographical unit or region in Fig. 1 to the *target-region*'s scope definition differs. For example in Fig. 2, *Amsterdam*, *Rotterdam* and *The Hague* contribute more to the definition of the *Netherlands* than other cities in the *Netherlands.* This contribution represents the importance attached to an entity in resolving the *target-region*'s geographical scope. Through experiment or expert knowledge a weight is attached to each geographical entity in Fig. 1.

The term *city* as used here has a broader sense referring to any of the following municipalities: cities, towns or villages. Fig. 2 shows the reference scope model for *the Netherlands* demonstrating how the geographical scope modelling data structure in Fig. 1 is populated. To determine the geographical scope of a document, terms that refer to locations such as place names and adjectives (referring to location and people) are extracted from the document. The extracted information is mapped against the reference geographical scopes of places, and a weighted list of geographical scopes associated with the document retrieved.

A number of algorithms can be used to implement the mapping from documents to reference geographical scopes modelled according to Fig. 1. The algorithms could be based on machine learning where the learner is trained with reference geographical scopes, and used to assign geographical scopes to new unseen documents. Alternatively, information retrieval algorithms such as vector space models could be used to index the reference geographical scope data. To assign scopes to documents, the search query consisting of place names ex-

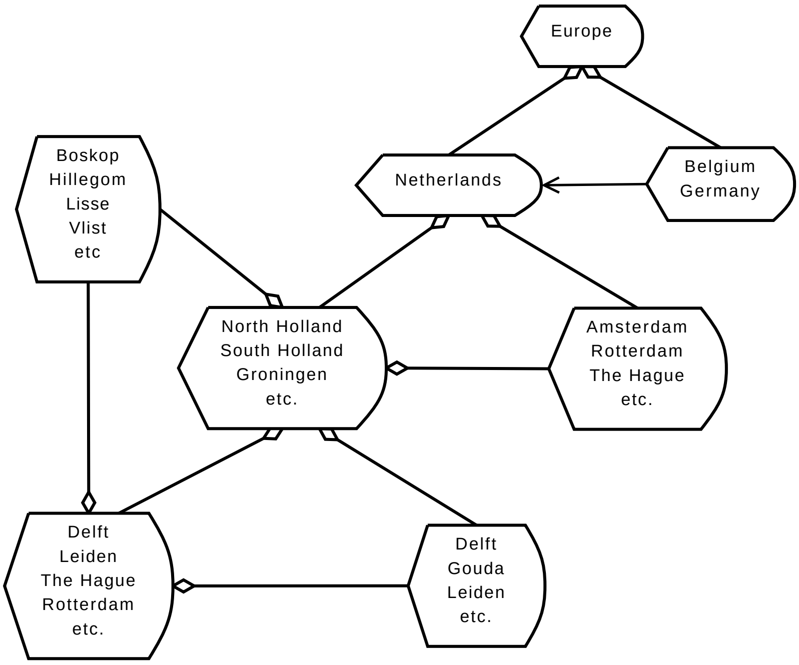

Figure 2: Sample data model for scope of the Netherlands.

tracted from the document might be run against the index, and a list of ranked scopes retrieved. The retrieved scopes would then be considered to represent the geographical coverage of the document. In this paper, we report on an information retrieval based algorithm to implement the mapping from document to reference geographical scopes.

The *zone indexing* [19] paradigm provides a suitable path to map documents to reference geographical scopes. The zone indexing model sub-divides documents into zones, and creates inverted indexes for each zone. The model supports querying against individual zones. Each zone can be assigned a weight reflecting its importance in the document. These weights are either assigned by an expert or through experiment. The score of each zone for a given query is combined to form the document score for the query.

The matching score between document d and geographical scope g is computed as:

$$score(g,\ d) = \sum_{j\,=\,1}^{n} \sum_{i\,=\,1}^{z} s(j)\ \times\ w_i(j) \tag{1}$$

where, n is the number of place names in the document, z the number of geographical zones, $w_i(j)$ the weight of zone i containing place name j, and $s(j) = 1$ if a match occurs between a document's and zone's place name, otherwise, $s(j) = 0$.

Zone	Data	Weight
Target region	Netherlands	0.30
Parent region	Europe	0.05
Neighbours	Belgium, Germany	0.03
Primary cities	Amsterdam, Rotterdam, The Hague	0.20
Child regions	Zeeland, Utrecht, Groningen	0.15
Child-child regions	Delft, Leiden, The Hague, Rotterdam	0.07
Secondary cities	Delft, Gouda, Leiden	0.15
Smallest cities	Boskop, Hillegom, Lisse, Vlist	0.05

Table 3: Zone index for the sample Netherlands scope data in Fig. 2.

To illustrate, the sample scope data for the *Netherlands* in Fig. 2 is transformed to the zone indexing paradigm as shown in Table 3. On the basis of the zone indexing, the geographical scope information in Fig. 2 is modelled as a document, and standard queries can be issued against the index to retrieve a ranked list of scope documents. Casting geographical scopes as documents is one of the novel contributions of this paper. The retrieved scope documents represent the geographical scopes of the query. Each query is made-up of place names extracted from documents whose scopes are to be resolved. The weights are assigned to each zone by intuition. For example, a geographical reference in the query which is also the name of a scope is accorded more weight for a match in a *target-region* zone than a match in a *parent-region* zone.

To demonstrate the use of Eq. 1, consider documents d_1 and d_2 with the list of toponyms they contain:

$$\begin{aligned} d_1 &= \{ \text{ Netherlands, Groningen, Leiden } \} \\ d_2 &= \{ \text{ Groningen, Leiden, Lisse } \} \end{aligned}$$

Using the sample reference scope for the *Netherlands* shown in Table 3, the documents, d_1 and d_2 are resolved to the *Netherlands* as follows:

$$score(g, d_1) = 0.30 \times 1 + 0.15 \times 1 + 0.07 \times 1 + 0.15 \times 1 = 0.67$$

$$score(g, d_2) = 0.15 \times 1 + 0.15 \times 1 + 0.07 \times 1 + 0.05 \times 1 = 0.42$$

The score formula ranks document d_1 higher than d_2 in the scope of *Netherlands*, i.e., document d_1 is considered more $Netherlands$ than document d_2. The Boolean based score ignores the number of times a term occurs in a document. This results in loss of frequency information which is important to guess to what degree a document is about a given term in the query. To include frequency count information, $s(j)$ is replaced with the frequency count $f(j)$ of the place name j in the document. Then Eq. 1 becomes:

$$score(g,\ d) = \sum_{j\,=\,1}^{n} \sum_{i\,=\,1}^{z} f(j)\ \times\ w_i(j) \tag{2}$$

where, $f(j)$ is the frequency count of the place name j in the document. Furthermore, places found in the same region or zone are not always of equal importance, e.g., in the *Netherlands*, *The Hague* can be considered more important than *Rotterdam* because it is the seat of the national government. Therefore, each place is assigned an importance weight $p_i(j)$ based on its type or population size or economy, and score formula in Eq. 2 becomes:

$$score(g,\ d) = \sum_{j\ =\ 1}^{n} \sum_{i\ =\ 1}^{z} f(j)\ \times\ w_i(j)\ \times\ p_i(j) \tag{3}$$

where, $p_i(j)$ is the importance attached to place j in zone i. To demonstrate consider document d_3 and d_4 with the list of place names they contain as follows:

$d_3\ =\ \{$ The Hague $\}$
$d_4\ =\ \{$ Rotterdam $\}$

Assume that places of type *capital city* are considered more important than other places, and therefore, are given an importance weight of 2.0, and to the rest an importance weight of 1.0. Using Eq. 3, documents d_3 and d_4 are resolved to the *Netherlands* with different scores

$$score(g, d_3) = 1\ \times (0.20\ \times\ 2.0\ + 0.07\ \times\ 1.0) = 0.47$$

$$score(g, d_4) = 1\ \times (0.20\ \times\ 1.0\ + 0.07\ \times\ 1.0) = 0.27$$

Document d_3 is more about the *Netherlands* than document d_4 though the place names in them share zones.

3.2. Using Anthroponyms

The *anthroponym-based* scope resolution strategy is grounded on Assumption 2.

Assumption 2. *VIPs (e.g., political leaders) in the same geographical region or at the same leadership hierarchy level tend to be mentioned together in a unit of a discourse. That is, presidents are most likely to be mentioned together with the members of their administration or with presidents of other countries in a unit of a discourse. For example, US President Barack Obama is most likely to be mentioned in a discourse together with US Vice President Joe Biden or President Yoweri Kaguta Museveni of Uganda in a discourse.*

In this paper, we shall call the people whose actions and opinions most determine course of events in a particular geographical scope or area *GeoVIP*s. The political and government leaders fall under the GeoVIP category because their jurisdictions are geographically constrained. It is therefore plausible to infer the geographical scope of a document from the list of GeoVIPs mentioned in a document. The people in other categories (e.g., builders and financial titans, artists and entertainers, heroes and icons, and scientists and thinkers) influence course of events more at global scope.

To effectively map GeoVIPs to the geographical area of their jurisdictions, the administrative division of a country can be used as a reference. Generally a country is sub-divided into two main administrative divisions: (1) the primary administrative divisions, and (2) the secondary administrative divisions. At every level of administrative division we have a GeoVIP who exercises an administrative jurisdiction over the area. As whatever happens in our neighbourhood affects us in one way or another, GeoVIPs can influence course of events in nearby geographical areas outside their jurisdictions. For example, President Barack Obama of United States of America (U.S.A.) can to some extent influence the course of events in countries bordering U.S.A. (i.e., in Canada and Mexico).

The names of people are highly ambiguous [36, 37] and therefore, exploiting person names to implement automatic systems is a non-trivial task. Before an attempt is made to resolve the geographical scope of documents using the names of people they contain, some minimum level of person name resolution is required. For example, we should be able to resolve the name *Mr. Cameron* to *Prime Minister David Cameron* before attempting to resolve the geographical scope *Mr. Cameron* indicates.

Generally the names of people are broadly grouped into three categories - first name, surname and middle name. Over the centuries people have shared names within these categories. The names of people can be shared within a given locality or/and names can be shared at global scale.

To perform automatic geographical scope resolution using the names of GeoVIPs, the names of people are grouped into three: (1) global names, (2) local or country-level names and (3) GeoVIP names. The global names category takes care of the frequency of sharing names globally (e.g., John, Paul, Joseph), and local or country-level names category are commonly found within a given geographic scope (e.g., Eriksson, Museveni, Kikwete, etc.).

We prefer names which are GeoVIP names, and are less frequently shared across the three groupings (i.e., globally, locally and GeoVIP). We exploit this bias to compute the weights assigned to each name according to Eq. 4:

$$NWF(name) = \sum_{i \subset \{g,\, l,\, v\}} K_i \log \frac{N_i + N_{max}}{N_i} \tag{4}$$

where NWF is the name weight factor, g stands for global category, l stands for local category, v stands for GeoVIP category, K_i category factor, N_i the number of persons sharing the name in category i and N_{max} the number of persons sharing the most common name globally. Category factor K_i weights the perceived importance of a category, e.g., names found in global names list are given less importance than names found in local names list. The expression in Eq. 4 computes an inverse of frequency, which is smoothed by using the logarithm.

To model each scope with GeoVIP information, GeoVIPs are grouped into five levels:

- VIP1 - names of GeoVIPs at the top-most hierarchy, e.g., for a country level scope, the top-most GeoVIP in the hierarchy is the president of the country.

- VIP2 - names of GeoVIPs next in hierarchy to GeoVIPs listed in VIP1 field. For example, at a country level scope, the cabinet ministers are included in this group.
- VIP3 - names of GeoVIPs next in hierarchy to GeoVIPs listed in VIP2 field. For example, at a country level scope, the members of the Parliament or the Senate are included in this group.
- VIP4 - names of GeoVIPs of neighbouring administrative units. For example, at a country level scope, GeoVIPs of the immediate neighbouring countries are listed.
- VIP5 - names of GeoVIPs of non-neighbouring administrative units. For example, at a country level scope, GeoVIPs of non-neighbouring countries are listed. This category caters for the assumption that GeoVIPs at the same level tend to be mentioned together in news stories. That is, it is more likely that *President Obama*, *Prime Minister Netanyahu* and *President Abbas* will be mentioned in the same story on *Middle East*.

An example GeoVIP grouping for Canada and United States of America is shown in Fig. 3. From Fig. 3 we can conclude that it is plausible to use GeoVIPs exercising jurisdictions at various administrative units or levels as evidence or clue to perform geographical scope resolution. The CIA[5] provides up-to-date lists of *Chiefs of State* and *Cabinet Members of Foreign Governments*. The *CIA World Leaders* list can be used to populate the anthroponym model to detect scopes of documents at the country level. Information about GeoVIPs below the level of *Cabinet Members* can be obtained from other sources such as the parliament, the senate, etc. However, we note that the challenges to using the GeoVIP list are:

1. Name ambiguity, e.g., many people share names locally and globally.
2. GeoVIPs serve limited terms in office, therefore, maintenance of a comprehensive, up-to-date GeoVIP information is non-trivial.

Similar to the *zone indexing* concept applied to *toponym-based* (see Section 3.1), the *anthroponym-based* approach uses the zone indexing strategy by modifying the formula in Eq. 3 as follows:

$$score(g,\ d) = \sum_{j}^{n} \sum_{i}^{z} f(j)\ \times\ w_i(j)\ \times\ m_i(j) \tag{5}$$

where $w_i(j)$ is the weight of the i^{th} VIP grouping where a match occurs for the name j, $f(j)$ is the frequency count of the name j in the document, $m_i(j)$

[5] `https://www.cia.gov/library/publications/world-leaders-1/index.html` [Accessed on 24 June 2010]

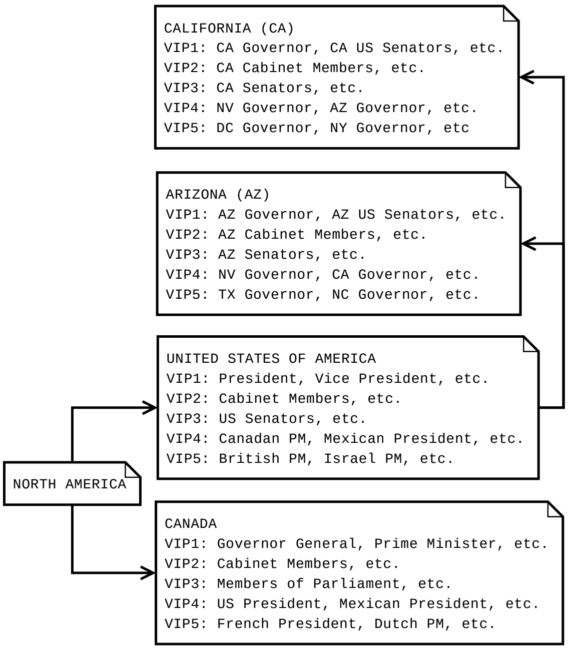
CALIFORNIA (CA)
VIP1: CA Governor, CA US Senators, etc.
VIP2: CA Cabinet Members, etc.
VIP3: CA Senators, etc.
VIP4: NV Governor, AZ Governor, etc.
VIP5: DC Governor, NY Governor, etc
ARIZONA (AZ)
VIP1: AZ Governor, AZ US Senators, etc.
VIP2: AZ Cabinet Members, etc.
VIP3: AZ Senators, etc.
VIP4: NV Governor, CA Governor, etc.
VIP5: TX Governor, NC Governor, etc.
UNITED STATES OF AMERICA
VIP1: President, Vice President, etc.
VIP2: Cabinet Members, etc.
VIP3: US Senators, etc.
VIP4: Canadan PM, Mexican President, etc.
VIP5: British PM, Israel PM, etc.
NORTH AMERICA
CANADA
VIP1: Governor General, Prime Minister, etc.
VIP2: Cabinet Members, etc.
VIP3: Members of Parliament, etc.
VIP4: US President, Mexican President, etc.
VIP5: French President, Dutch PM, etc.

Figure 3: Example U.S.A & Canada GeoVIP grouping.

is the weight of the name j with respect to the i^{th} grouping computed using Eq. 4. To achieve a better result, the first and last names or a mention of the office of the target VIP candidate must appear at least once somewhere in the document (e.g., both the names of the US President must be mentioned at least once *Barack Obama* possibly with the title *President* or *US President*).

3.3. Scope Resolution Evaluation

This section describes tests run to validate the performance of the proposed scope resolution strategies against human annotated dataset. To measure the effectiveness of scope resolution systems in a standardized fashion, two things are needed:

1. A gold standard dataset consisting of a reference gazetteer and a reference document collection with each document marked with geographical scopes.
2. An evaluation metric to assess the correctness of system assigned scopes against gold standard scopes.

The current state-of-the-art approach to evaluate scope resolution systems uses a binary metric[38]. The binary metric in Eq. 6 is used to measure the performance of scope resolution strategies described in this paper. It assigns a score of 1 to a document when all the document's n scopes are listed at the top n ranked positions, otherwise it assigns 0. No attention is paid to correct document scopes ranked outside the top n ranked positions.

No. Scopes	One	Two	Three	Four+	Total
CoNLL-2003	702	318	73	31	1,124
$GS(\%)$	94.00	64.47	26.03	16.13	79.09

Table 4: Toponym-based scope resolution result.

No. Scopes	One	Two	Three	Four+	Total
News Articles	37	5	1	0	43
$GS(\%)$	68.0	0.0	0.0	–	58.0

Table 5: Anthroponym-based scope resolution result.

$$GS = \frac{\mid documents\, with\, correctly\, assigned\, scopes \mid}{\mid documents\, with\, scopes\, in\, the\, collection \mid} \tag{6}$$

3.3.1. Toponym-based Evaluation

The *toponym-based* approach is evaluated on the CoNLL 2003 English dataset [1]. The CoNLL 2003 English dataset is derived from the Reuters English Corpus (RCV1) [39]. The CoNLL 2003 Shared Task training and development dataset consists of 1,162 English language documents. Of the 1,162 documents, 1124 documents contain geographical terms, i.e., place names and geographical adjectives. The documents are assigned geographic scopes at the country level. Of 1,124 documents 702 are assigned single scopes, 318 double, 73 triple and 31 four or more scopes. The 1,124 documents share 514 unique names and 143 unique scopes with each document having 2.5 place names on average.

The scope resolution procedure described in this paper can assign scope up to six levels: continent, continent-directional, country, country-directional, province and province-directional. For this evaluation, the country level resolution is turned on as the geographical scopes assigned in the CoNLL collection are at country level. The system assigns multiple scopes to each document ranking from the most relevant to least relevant. Table 4 shows the summary of system performance computed using Eq. 6. The overall system performance is very good for documents with one (i.e., recall of 94%) and two scopes (i.e., recall of 64%), but very poor for documents with three or more scopes. This is comparable to the recall value of 95% on documents with single scopes reported in [38].

3.3.2. Anthroponym-based Evaluation

The *anthroponym-based* approach is evaluated on news stories collected from Ugandan news websites. The collection consists of 43 documents with a total of 25 Ugandan scopes at district levels with a total of 167 Ugandan VIP names.

The performance of anthroponym-based scope resolution with GeoVIPs on news articles is shown in Table 5. Overall it shows poorer performance in comparison to the toponym-based strategy. This experiment shows that exploitation

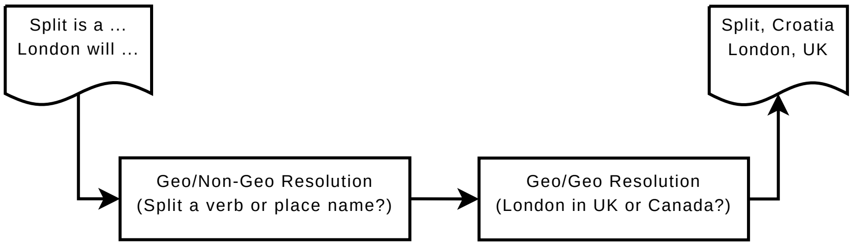

Figure 4: Toponym resolution schematic.

of VIP names found in documents could help in resolving geographical scopes of documents especially when the documents contain no mention of places.

4. Toponym Resolution

The names of places are ambiguous in many ways. They can reference other named entities (e.g., names of people, names of organizations, etc.), and may be used as common language vocabulary words (e.g., *Split* is a city in Croatia, *Over* a city in Germany, etc.). The use of place names outside their geographical or location context is sometimes referred to as *Geo/Non-Geo* ambiguity [6]. Besides reference or geo/non-geo ambiguity issues, names of places are referentially ambiguous. Referential ambiguity occurs when a place name references multiple places. This is also termed as *Geo/Geo* ambiguity [6]. And as we noted above, places are also referred to by more than one name, e.g., Netherlands vs. Holland.

The task of *toponym resolution* [2] is to map a place name to a non-ambiguous location or a geographical point on the surface of the Earth. This mapping is normally done using a geographical reference coordinate system such as latitude and longitude. The terms geographical name, place name and toponym are synonymous, and they are used interchangeably to mean the same thing (i.e., a name of a place) throughout this paper.

4.1. Toponym Resolution Procedure

Figure 4 shows a schematic diagram depicting the toponym resolution process. In this work geo/non-geo ambiguity resolution is performed with the help of an off-the-shelf named entity recognition tool, the Alias-i LingPipe. [6] The motivation is that place name recognition components of the state-of-the-art named entity recognizers have achieved near human performance [1], so that, existing off-the-shelf recognizers are sufficient to perform geo/non-geo resolution task.

The toponym approach reported in this paper exploits $26,820$ geographical scopes automatically assigned to documents (using scope resolver described

[6] `http://alias-i.com/lingpipe/` [Accessed on 08 June 2010]

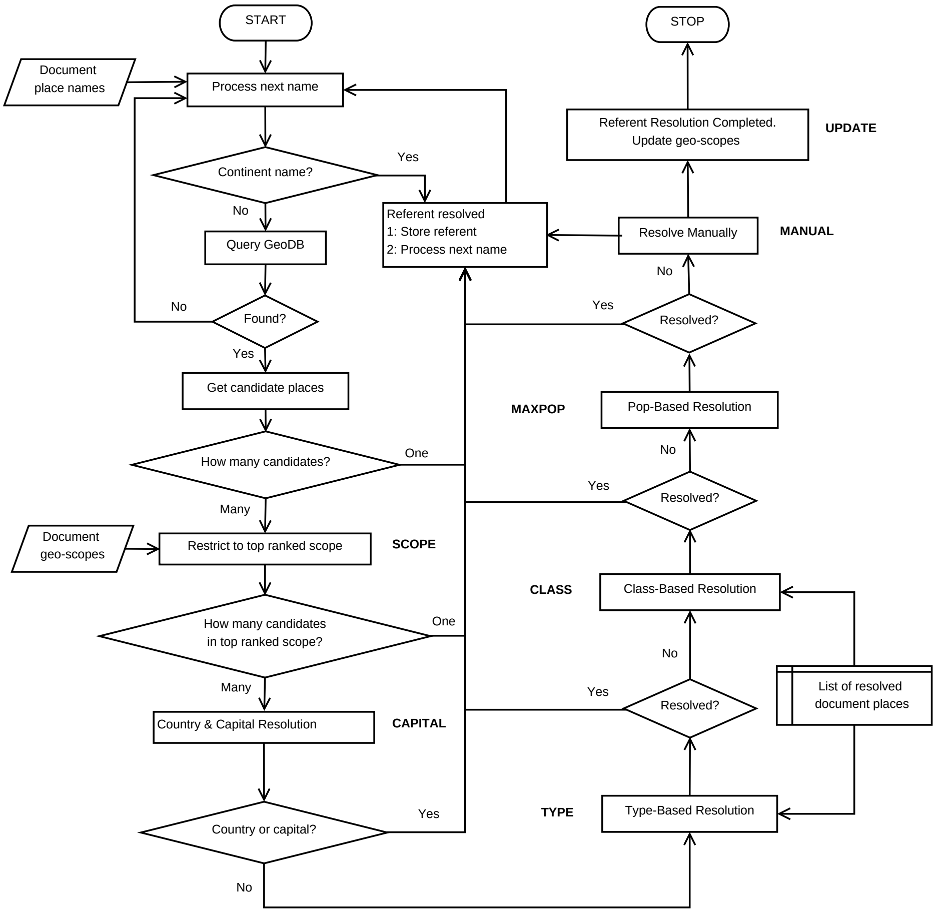

Figure 5: Toponym resolution algorithm. Document scopes obtained from global analysis are used here (left side, half-way down) to inform detailed toponym resolution questions.

in Sec. 3), the type of place (e.g., city), the classification of place (e.g., populated place, administrative division, etc.), the population of the place, and the frequency of non-ambiguous or resolved places.

This work proposes a *geographical scope-driven* toponym resolution approach. The approach builds on previous work discussed in Sec 2. The approach reported here applies the following techniques at various stages of resolution:

1. Single referent per discourse;
2. Scope restriction;
3. Population heuristics;
4. Place type restriction;
5. Default sense heuristics.

Figure 5 depicts the proposed toponym resolution routines. We note that the scope restriction and place type heuristics as applied in this paper are novel.

We first present an explanation of how the various blocks work together to accomplish the resolution task.

4.1.1. Single Referent and Default Sense Heuristics

The toponym resolution algorithm starts by invoking the single referent per discourse and default sense heuristics. Place names mentioned more than once in a document are assumed to refer to the same place on the basis of the *single referent per discourse* heuristic. Next the *default sense* heuristic is invoked to assign place names with continent sense to the continents. Candidate places for toponyms with senses other than continent senses are obtained from the geographical database (GeoDB). Toponyms with single candidate places are resolved to these places, and toponyms with multiple candidates are passed to lower processing modules starting with the scope restriction module. However, place names with no candidates in the GeoDB are ignored. This often happens when a geographic database lacks complete coverage.

4.1.2. Scope Restriction Heuristic

The scope restriction heuristic (see SCOPE in Fig. 5) extends the country level restriction reported by [4]. The heuristic exploits an elaborate list of ranked geographic scopes assigned to a document by the geographic scope resolver described in Sec. 3. A toponym with multiple candidate referents is assigned to a single most highly ranked document geographical scope. The other candidates belonging to lower ranked document scopes are discarded. If a selected scope contains a single candidate, the candidate is marked as the place referred to by the name. However, if a selected scope contains multiple candidates with the same name, it is passed to the next processing block in the hierarchy, i.e., country and capitals resolution heuristic (see CAPITAL in Fig. 5). The accuracy of the scope restriction heuristic depends on how well the scope resolution performed. The greater the error in scope resolution the greater the error in referent resolution.

4.1.3. Country and Capital Heuristic

The country and capitals heuristic (see CAPITAL in Fig. 5) is a kind of default sense heuristic, but restricted within the selected geographical scope for the name. A toponym's candidate place of type country or national capital or provincial capital is selected as the place being referred to within the selected scope. The order of preference is: *country* $\mapsto$ *country capital* $\mapsto$ *provincial capital*.

4.1.4. Type and Class Co-existence

The type-based heuristic (see TYPE in Fig. 5) exploits the types of resolved places as a basis to resolve among competing candidate places. Commonly occurring types are preferred. The assumption is that places of the same type are more likely to be mentioned in a discourse. The candidate place of the type matching the most commonly occurring type among the resolved places is selected as the place being referenced in the text.

On the other hand, the class-based heuristic (see CLASS in Fig. 5) exploits geographic feature classifications in the Geonames.org[7] database. The `Geonames.org` categorises geographic features into nine broad classes:

1. Administrative unit (i.e., country, state, region, etc.).
2. Hydrographic (i.e., stream, lake, bay, etc.).
3. Locality or area (i.e., parks, area, nature reserve, etc.).
4. Populated place (i.e., city, town, village, etc.).
5. Road or railroad (i.e., road, railroad, tunnel, etc.).
6. Spot (i.e., spot, building, farm, etc.).
7. Hypsographic (i.e., mountain, hill, island, etc.).
8. Undersea (i.e., basin, undersea, range, etc.).
9. Vegetation (i.e., forest, heath, pine grove, etc.).

The class-based heuristic procedure is similar to the type-based heuristic. Similar to the type-based heuristic assumption, the places of a similar class are more likely to be mentioned in a discourse. And therefore, this heuristic selects the candidate place of the class matching the most frequently occurring class among the resolved places as the referenced place.

4.1.5. Population, Manual and Scope Update

The population-based heuristic (see MAXPOP in Fig. 5) is straightforward in that the place with the largest population is selected as the place being referred to. However, this heuristic is applied to candidate places of the same type and class, e.g., the town of *Groningen* in *Germany* with population of 4,166, and the town of *Groningen* in *Suriname* with population of 3,216. The population-based heuristic can be effective only when the population information of population centres are complete in the geographical database.

The manual resolution (see MANUAL in Fig. 5) is the last option when all the previous automated procedures fail to solve a given ambiguity problem. The task is passed over to the user to decide the meaning of the remaining ambiguous places by exploiting other sources of information at her or his disposal.

Because the toponym resolution component is part of the *place ambiguity resolution* system, its output is also used improve the accuracy of the *scope resolution* component. Therefore, upon completion of toponym resolution (see UPDATE in Fig. 5), the list of geographical scopes is updated by including scopes containing resolved places and their ancestor scopes. The other remaining scopes in the original ranked list are discarded.

4.2. Toponym Resolution Evaluation

The toponym resolution scheme proposed in this paper is evaluated on the TR-CoNLL datasets [2], and the *precision*, *recall* and *f-score* measures are used in the evaluation. Naturally our evaluation ignores manual resolution. Following the straightforward instantiation of the standard definition from [2],

[7] `http://www.geonames.org/about.html` [Accessed on 03 October 2009]

Precision, P is the ratio of the number of correctly resolved toponym instances, T_C and the number of toponym instances that the system attempted to resolve (either correctly, T_C or incorrectly, T_I)

$$P = \frac{T_C}{T_C + T_I} \tag{7}$$

Recall, R is the ratio of the number of correctly resolved toponym instances, T_C and the number of all toponym instances, T_N (i.e., the number of resolvable toponyms in a text document or corpus)

$$R = \frac{T_C}{T_N} \tag{8}$$

Note that $T_N = T_C + T_I + T_U$ where T_U is the number of toponym occurrences whose candidate referents are unresolved.
F-Score, F_β for precision P and recall R is defined as

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{9}$$

and the *F-Score* at $\beta = 1$ is

$$F_{\beta=1} = \frac{2PR}{P + R} \tag{10}$$

and all the *F-Score* values reported here are computed at $\beta = 1$.

The toponyms in TR-CoNLL corpus are annotated with a spatial footprint (i.e., latitude and longitude) information from the TextGIS [2]. On the other hand, the scope resolution strategies and referent resolution routines proposed in this work use the Geonames.org [8] database as the source of their spatial information. The difference in geographical information results in spatial information mismatch at evaluation. Any two points in TextGIS and Geonames.org are assumed similar if and only if they are separated by not more than 0.04° (i.e., $\approx$ 4.48 Kilometres) in both latitude and longitude direction. Therefore, *Seoul* at $< 37.5664; 127.0 >$ and *Seoul* at $< 37.5663889; 126.9997222 >$ are assumed to be the same *Seoul* from the two databases.

The *LSW03* system in [2] is grounded on two minimality heuristics: (1) the *one-referent-per-discourse* heuristic that assumes that a place name mentioned in a discourse refers to the same location throughout the discourse, and (2) the *spatial minimality* heuristic that assumes that, in cases where there is more than one toponym mentioned in some span of text, the smallest region that is able to ground the whole set is the one that gives them their interpretation. The *RAND* and *MAXPOP* formed the baseline heuristics in [2]. The *RAND* heuristic selects a random referent if at least one referent was found in the gazetteer, and the *MAXPOP* picks the candidate referent with the largest population. Table 6

	Heuristic	Precision	Recall	F-Score
Leidner [2]	RAND	0.2973	0.2973	0.2973
	MAXPOP	0.6506	0.1976	0.3032
	LSW03	0.3650	0.3177	0.3397
Mahali	MAXPOP	0.6829	0.2864	0.4035
	CAPITAL	0.6305	0.5326	0.5774
	SCOPE	0.5230	0.4409	0.4785
	SCOPE+CAPITAL	0.7529	0.6359	0.6895
	SCO+CAP+TYP+CLA	0.7744	0.6541	0.7092
	ALL	0.7754	0.6549	0.7101

Table 6: Toponym resolution results on TR-CoNLL. MAXPOP and CAPITAL are baselines; SCOPE uses scope restriction; CAPITAL uses countries and capitals; SCO+CAP+TYP+CLA is combination of SCOPE, CAPITAL, TYPE and CLASS; and ALL is combination all the models in Fig. 5. See text for details.

shows the performance of the various models of the new proposed toponym resolution strategy in comparison to *RAND, MAXPOP* and *LSW03* from [2].

The top block in Table 6 shows the performance data from [2] on the TR-CoNLL corpus, and the bottom block shows the performance of Mahali [9] toponym resolution component on the TR-CoNLL corpus. The heuristics *CAPITAL* and *MAXPOP* are treated as the baseline schemes. CAPITAL selects the candidate location that refers to *country*, *national capital* or *provincial capital* as the place referred to. The MAXPOP heuristic picks the candidate location that has the largest population as the place referred to. For schemes *SCOPE* (scope restriction), *SCOPE+CAPITAL* (i.e, combination of SCOPE and CAPITAL), *SCO+CAP+TYP+CLA* (i.e, combination of SCOPE, CAPITAL, TYPE and CLASS) and *ALL* (i.e., when all the models are activated) refer to Fig. 5. The baseline heuristic CAPITAL is very competitive on the TR-CoNLL corpus. This reflects the types of toponyms used in stories with global scopes (i.e., countries and capital cities are commonly mentioned in stories with global scopes). The TR-CoNLL corpus consists of stories with scopes of the global, and therefore, a good performance is expected in terms of recall with CAPITAL heuristic. The 10% performance improvement of our MAXPOP heuristics over the MAXPOP heuristic in [2] can be attributed to the difference in geographical database used in the two implementations. The SCOPE heuristic which selects candidate locations found in the top ranked scopes shows a good performance as well. As more heuristics are combined the performance improves smoothly as seen in SCOPE+CAPITAL, SCO+CAP+TYP+CLA and ALL. Overall the scheme proposed in this work shows that it is very competitive.

[8] `http://www.geonames.org` [Accessed on 24 June 2010]

[9] All the components developed in the course of this work form part of the system called *Mahali*. *Mahali* means ‘place’ in Kiswahili.

The superiority of our approach lays in the fact that we compute toponym resolution within the document's scopes. As long as the scopes of documents are of highest quality, the document's toponyms stand a good chance to be resolved accurately to places they refer to. We argue that it is a good premise to first detect the scopes of documents before computing the whereabouts of geographical entities located in the document.

5. Query Expansion

The motivation for query expansion is to reduce the mismatch between query and document by expanding the query terms using words or phrases which are synonymous to query terms or share other semantic/conceptual relationships with the terms contained in the set of relevant documents. Incorporating geographical information or metadata in a query modification component and ranking algorithm could positively influence results returned against geography-dependent user needs. This section explores query expansion strategies for a geographically constrained information retrieval task. The query expansion approach reported here investigates the application of relevance feedback (i.e., blind and explicit feedback) procedures to improve retrieval by adding toponyms found in the relevant documents to original query. Two relevance feedback schemes are explored – one approach adds toponyms found in the relevant documents directly, and the other adds toponyms found in geographical scopes of relevant documents. We shall call the first approach the *toponym-based* scheme and the second the *scope-based* scheme.

5.1. Toponym-based Expansion

The toponym-based expansion approach adds toponyms derived from relevant documents [10] to the original query with the view of tilting search results towards documents within the user's geographical region of interest. The new query vector of toponyms is formulated as

$$\overrightarrow{g}_{new} = \overrightarrow{g}_{old} + \overrightarrow{g}_{rel} \tag{11}$$

where, $\overrightarrow{g}_{old}$ is the vector of toponyms in the original search query, g_{rel} is the vector of toponyms in the N relevant documents, and g_{new} is the new vector of toponyms to the new search query for relevant feedback. The formula in Eq. 11 is motivated by Rocchio's feedback formula [see 19, Chapter 9] shown below.

$$\overrightarrow{q}_m = \alpha\overrightarrow{q}_o + \beta\frac{1}{|D_r|}\sum_{\overrightarrow{d}_j\in D_r}\overrightarrow{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\overrightarrow{d}_j\in D_{nr}}\overrightarrow{d}_j \tag{12}$$

where, $\overrightarrow{q}_o$ is the original query vector, D_r and D_{nr} are the set of known relevant and non-relevant documents retrieved by q_o respectively, and α, β, and γ are

[10] Relevant documents are derived through blind feedback or explicit feedback procedure.

weights attached to each term. We restrict our attention to the geographical terms in the vectors of $\overrightarrow{d}_j$.

The relevant toponym vector can be represented using its member elements as follows

$$\overrightarrow{g}_{rel} \equiv \langle f_1, f_2, \ldots, f_n \rangle \tag{13}$$

where, f_i is the i^{th} toponym in the top N relevant documents. The weights of the toponyms in $\overrightarrow{g}_{rel}$ are computed using

$$weight(f_i) \;=\; \frac{rel(f_i)}{rel(f_i)\;+\;non(f_i)} \tag{14}$$

where, $rel(f_i)$ is the number of occurrences of f_i in the relevant document set, and $non(f_i)$ is the number of occurrences of f_i in the non-relevant document set.

5.2. Scope-based Expansion

Scope-based query expansion exploits the geographical scopes assigned to documents to select toponyms to expand the original query. The geographical scope resolver used to tag document scopes is described in Sec 3, and achieved an overall score of 79.09% (see Table 4) on news article collections. We recall here that *geographical scope* is not a linguistic or textual notion, but rather a geographical one, roughly the geographical region relevant to a particular reference or document. The scope resolver assigns multiple weighted scopes to each document. The scope-based approach selects the frequently occurring scopes from the relevant documents to expand the set of search query's toponyms. The candidate toponyms to expand search queries are those belonging to the most frequent scopes shared by the most relevant documents.

The weight of place names added to original search query is computed by

$$weight(f_i) \;=\; \log\left(1 + \frac{|S(M_i)|}{|S|}\right) \tag{15}$$

where, S is a set consisting of the most frequent scopes shared among the N relevant documents, $|S(M_i)|$ is the number of scopes to which toponym M_i belongs in the scope set S, $|S|$ is the total number of scopes selected from the top N relevant documents.

5.3. Query Expansion Evaluation

The query expansion procedures reported here are evaluated on the GeoCLEF 2007 [40] dataset which provides 25 geographically focused topics. Experience has shown that the original search query terms should be preserved in the new feedback query formulation to achieve performance improvement [41]. The TREC evaluation tool [11] is used to evaluate retrieval results. We evaluate

[11] `http://trec.nist.gov/trec_eval/` [Accessed on 10 November 2009]

	Whole document set					Residual document set			
	LUD	LUB	LUE	TOP	SCO	LUD	LUE	TOP	SCO
Num of query	25	25	25	25	25	25	25	25	25
Num retrieved	25000	25000	25000	25000	25000	25000	25000	25000	25000
Num relevant	650	650	650	650	650	545	545	545	545
Retrieve relevant	603	613	583	613	612	510	486	511	515
MAP	0.2724	0.2858	0.3231	0.4245	0.3057	0.1850	0.1831	0.1524	0.2025
Relevant used	–	15	5	5	5	–	5	5	5
Recall	0.9277	0.9431	0.8969	0.9431	0.9415	0.9358	0.8917	0.9376	0.9450

Table 7: Summary of feedback evaluation on the whole and residual document sets.

query expansion on both the *whole document set* and *residual document set*. The whole document set include documents used to construct the new feedback query, while the residual document set consists of documents not used in construction of new feedback query. The motivation for evaluation on residual document set is that the relevant documents seen by the user in the previous search are irrelevant in the next search.

Table 7 shows the results of relevance feedback procedures on both the whole and residual document set. *LUD*, *LUB* and *LUE* are the default Lucene, Lucene with blind feedback and Lucene with explicit feedback runs respectively. *TOP* shows the result of the toponym-based approach with five (5) relevant and five (5) non-relevant documents. *SCO* shows the result of scope-based scheme using five (5) relevant documents for the query expansion procedure. The *SCO* chose the top 5 scopes to constrain place name selection for feedback query expansion.

The following observations can be made from the results:

1. The toponym-based (*TOP*) scheme shows superiority when evaluated on the whole set of document collection which includes documents used to construct the new search query for relevance feedback. The improvement is as a result of ranking highly the documents used to construct the new search query. When the scheme is evaluated against residual document collection it achieved the worst performance in comparison to the default Lucene and the scope-based (*SCO*) scheme.
2. The scope-based (*SCO*) scheme shows superiority when evaluated on the residual document collection set which consists of documents not used to construct the new search query for relevance feedback. Though the scheme performed poorly against *LUE* and *TOP* on the whole document collection, it performed better than all the other schemes on residual document set.
3. From the observations we can conclude that it is possible to achieve document retrieval accuracy improvement with a careful integration of geographical metadata such as toponyms into query expansion procedure.

6. Relevance Ranking

Relevance ranking is the task of ordering the retrieved set of documents by relevance to the user's information needs so that the most relevant documents are pushed to the top of the ranked result list. This section describes two types of relevance ranking schemes which exploit geographical scopes and feature types in documents and search queries to rank documents by geography. The *scope-based* metric is used to rank documents for queries which are resolvable to at least one scope. On the other hand, the *type-based* metric is used to rank documents when a query mentions only the geographical subjects, e.g., *lakes with monsters.* The scores of the non-geographic component and the geographical components are combined through linear interpolation and through weighted harmonic means.

6.1. Scope-based Metric

The scope-based relevance measure uses geographical scopes assigned to queries and documents to rank documents according to query geographic restrictions similar to schemes explored in [42]. The scope-based relevance measure is defined as:

$$ScopeSim(q,d) = \sum_{s} \sqrt{wt_{(q,s)}} \times log(1 + wt_{(d,s)}) \tag{16}$$

where $wt_{(q,s)}$ is the weight assigned to scope s in query q by the scope resolver, and $wt_{(d,s)}$ is the weight assigned to scope s in document d by the scope resolver. Eq. 16 attributes more importance to query scopes than document scopes. We are interested in satisfying the searcher's information need within the scopes the searcher implies in the query. Therefore, it is prudent to structure the scope-based similarity measure in a fashion that favours query scopes over document scopes.

6.2. Type-based Metric

The type based relevance measure utilizes the geographical feature class and type defined in a database of geographic features to compute a document's relevance to a query. The measure ranks documents by query feature type restriction. The feature class and type as defined in the Geonames.org[12] database are used to implement the type-based relevance measure. Types are lowest-level classifying groups, and classes are groups of types (see Fig. 6). Figure 6 shows the structure of the Geonames.org [13] feature grouping hierarchy, where, A is administrative unit, H is hydrographic, L is locality or area, P is populated place, R is road or railroad, S is spot, T is hypsographic, U is undersea and V is vegetation.

The type based relevance measure is defined as:

$$TypeSim(q,d) = \frac{1.0}{\sqrt{1 + \frac{N_{qFClass} - N_{qFType}}{N_{qFClass}}}} \tag{17}$$

[12] http://www.geonames.org/export/codes.html [Accessed on 10 June 2010]

[13] http://www.geonames.org [Accessed on 10 June 2010]

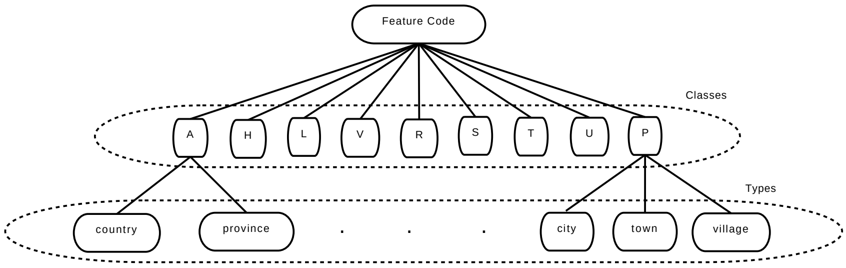

Figure 6: Sample Geonames.org feature code hierarchy.

where; $N_{qFClass}$ is the number of occurrences of the required query feature class in the document, and N_{qFType} is the number of occurrences of the required query feature type in the document. The maximum value of 1.0 is reached for Eq. 17 when the number of $N_{qFClass}$ is equal to N_{qFType}. This happens when all features of type $FType$ are of class $FClass$.

6.3. Combining Geographical and Non-Geographical Metrics

This section describes attempts that have been proposed to combine the non-geographical relevance measures and the geographical relevance measures to a unified relevance measures.

6.3.1. Linear Interpolation

The linear interpolated combination (LIC) is derived as:

$$\begin{aligned} LIC(q,d) &= \lambda_T\ NonSim(q,d) + \lambda_G\ GeoSim(q,d) \qquad (18) \\ \lambda_T\ +\ \lambda_G &= 1 \qquad (19) \end{aligned}$$

where; $NonSim(q,d)$ is the non-geographical metric, λ_T is the non-geographical interpolation factor (NIF) and λ_G is the geographical interpolation factor (GIF). The non-geographical and geographical scores are normalized to $[0,1]$ before linearly combining the ranked lists. The $GeoSim(q,d)$ in Eq. 18 is replaced by either Eq. 16 or Eq. 17 depending on the nature of the query.

6.3.2. Harmonic Mean Interpolation

The weighted harmonic mean (WHM) combination borrows from the classic precision and recall combination formula, the *F-measure* [see 43, Chapter 7] commonly used to measure performance of information retrieval (IR) systems. The motivation is to determine the importance of non-geographical relevance relative to geographical relevance, and then use the insight to rank documents by both non-geographical and geographical relevance. The weighted harmonic mean (WHM) combination is defined as:

$$WHM(q,d) = \frac{(1+\beta) \times GeoSim(q,d) \times NonSim(q,d)}{\beta \times GeoSim(q,d) + NonSim(q,d)} \qquad (20)$$

where; β indicates the importance attached to either $GeoSim(q,d)$ or $NonSim(q,d)$ in the unification. The following special cases are derived as a consequence of harmonic mean combination:

1. if $\beta = 1$, equal importance is attached to both non-geographical and geographical relevance.
2. if $\beta = 0$, no importance is attached to non-geographical relevance.
3. if $\beta = \infty$, no importance is attached to geographical relevance.

The interesting feature of this combination is that an optimal value of β where the best performance is achieved can be spotted. The $GeoSim(q,d)$ in Eq. 20 is replaced by either Eq. 16 or Eq. 17 depending on the nature of the query.

6.3.3. Extended harmonic mean combination

The extended harmonic mean (EHM) combination linearly adds the non-geographical relevance measure $NonSim(q,d)$ to the weighted harmonic mean (WHM) combination (see Eq. 20) as follows:

$$EHM(q,d) = NonSim(q,d) + \frac{(1+\beta) \times GeoSim(q,d) \times NonSim(q,d)}{\beta \times GeoSim(q,d) + NonSim(q,d)} \quad (21)$$

The $GeoSim(q,d)$ in Eq. 21 is replaced by either Eq. 16 or Eq. 17 depending on the nature of the query.

6.4. Evaluation

The proposed relevance measure and weighting schemes are evaluated on GeoCLEF 2007 [40] dataset. In [44] geographic topics are categorized into eight groups according to the way they depend on a place (e.g., Netherlands, Texas, etc), geographic subject (e.g., city, river, etc.) or geographic relation (e.g., north Groningen, western Europe, etc.). The ranking parameters in the formula for the experiment are tuned on the GeoCLEF 2006 [44] dataset (i.e., the topics and document collection). The ultimate purpose of the experiment is to compare the proposed relevance ranking schemes against the default search engine relevance ranking. Therefore, efforts were made to construct high quality queries to run against the document collections.

Table 8 shows the best five entries in GeoCLEF 2007 campaign [45] where there were eleven competing teams in total. LIC, WHM and EHM show the performance of relevance ranking formulas in Eq. 18, Eq. 20 and Eq. 21 respectively. The EHM performs slightly better than the best entry *catalunya* by a margin of 2.98%, and by a margin of 8.9% over Lucene.

We note that the best performance is achieved when the importance of non-geographical relevance outweighs the importance of geographical relevance by a large factor. [46] reported an improvement when geographical terms in the query are weighed half or less than the weight of non-geographical terms, which is in agreement with our observation. The difference is that we incorporate geographical scope information into relevance ranking algorithm while [46] implement term weighting in the context of query expansion procedure. With properly balanced contributions of the non-geographical and geographical scores to a combined score (see Eq. 21), an improvement can be achieved.

	Rank	Participant	MAP
GeoCLEF 2007 [40]	1^{st}	catalunga	28.50%
	2^{nd}	cheshire	26.42%
	3^{rd}	valencia	26.36%
	4^{th}	groningen	25.15%
	5^{th}	csusm	21.32%
Metrics in this paper.	1^{st}	Eq. 21: EHM ($\beta \geq 3.5$)	29.35%
	2^{nd}	Eq. 20: WHM ($\beta \geq 50$)	27.49%
	3^{rd}	Eq. 18: LIC ($\lambda_T = 0.9$)	27.10%
	4^{th}	$NonSim(q, d)$: Lucene	26.95%

Table 8: Comparison to GeoCLEF 2007 participants.

7. Conclusion & Discussion

This paper sets out to investigate the argument that geographical information contained in documents and search query texts may be useful to improve the quality of information retrieval especially when the user formulates queries in a geographical information retrieval (GIR) setting. New schemes were proposed to extract geographical information from documents, and the evaluation of these techniques yielded promising results with both toponym and scope resolution routines achieving performance accuracy of 70% and above. The query expansion and relevance ranking modification within the document scope framework yielded a retrieval accuracy improvement of 9% over the standard IR system. We have shown that if the searcher is patient enough to select relevant documents from the results of previous query, retrieval accuracy improvement can be achieved by expanding the query's geographical component with toponyms from the user-selected relevant documents. We have further shown that by integrating geographical scope information into relevance-ranking procedures, retrieval accuracy can be improved.

8. Acknowledgements

This work was supported by NUFFIC within the Framework of Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education and Capacity (NPT) under project NPT/UGA/132 on "Building a sustainable ICT training capacity in the public universities in Uganda".

References

[1] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language independent named entity recognition, in: Walter Daelemans and Miles Osborne, Editors, Proceedings of CoNLL-2003, 2003, pp. 142–147.

[2] J. L. Leidner, Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh (2007).

[3] B. Martins, N. Cardoso, M. Chaves, L. Andrade, M. J. Silva, The University of Lisbon at GeoCLEF 2006, in: Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 20-22, 2006, Revised Selected Papers, Vol. 4730/2007, Springer (Lecture Notes in Computer Science LNCS), 2007, pp. 986–994.

[4] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fluart, W. Zaghouani, A. Widiger, A.-C. Forslund, C. Best, Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation, in: Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC), 2006, pp. 53–58.

[5] B. Martins, M. J. Silva, A Graph-Ranking Algorithm for Geo-Referencing Documents, in: Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining, Texas, USA, 2005.

[6] E. Amitay, N. Har'El, R. Sivan, A. Soffer, Web-a-Where: Geotagging Web Content, in: Proceedings of SIGIR-04, the 27th Conference on Research and Development in Information Retrieval, Sheffield, South Yorkshire, UK, 2004.

[7] W. Zong, D. Wu, A. Sun, E.-P. Lim, D. H.-L. Goh, On Assigning Place Names to Geography Related Web Pages, in: Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries, 2005, pp. 354–362.

[8] J. Ding, L. Gravano, N. Shivakumar, Computing Geographical Scopes of Web Resources, in: Proceedings of the 26th Very Large Data Bases (VLDB) Conference, Morgan Kaufmann Publishers Inc., 2000, pp. 545–556.

[9] C. E. C. Campelo, C. de Souza Baptista, Geographic scope modeling for web documents, in: Proceedings of Workshop on Geographic Information Retrieval (GIR'08), Napa Valley, California, USA, 2008.

[10] I. Anastácio, B. Martins, P. Calado, A Comparison of Different Approaches for Assigning Geographic Scopes to Documents, in: Proceedings of the 1st INForum-Simpósio de Informática, 2009.

[11] E. Rauch, M. Bukatin, K. Baker, A Confidence-Based Framework for Disambiguating Geographic Terms, in: Kornai, A. and Sundheim, B. (eds) Proceedings of the HTL-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada, 2003, pp. 50–54.

[12] B. Pouliquen, R. Steinberger, C. Ignat, T. D. Groeve, Geographical Information Recognition and Visualisation in Texts Written in Various Languages, in: Proceedings of ACM (SAC2004), Nicosia, Cyprus, 2004.

[13] P. Clough, Extracting Metadata for Spatially-Aware Information Retrieval on the Internet, in: Proceedings of Workshop on Geographic Information Retrieval (GIR'05), CIKM2005, Bremen, Germany, 2005.

[14] B. Martins, M. J. Silva, S. Freitas, A. P. Afonso, Handling Locations in Search Engine Queries, in: Proceedings of the 3rd Workshop on Geographic Information Retrieval held at The 29th Annual International ACM SIGIR Conference, Seattle, WA, USA, 2006.

[15] R. Volz, J. Kleb, W. Mueller, Towards ontology-based disambiguation of geographical identifiers, in: Proceedings of WWW2007, 2007.

[16] D. A. Smith, G. S. Mann, Bootstrapping Toponym Classifiers, in: Kornai, A. and Sundheim, B. (eds) Proceedings of the HTL-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada, 2003, pp. 45–49.

[17] J. L. Leidner, G. Sinclair, B. Webber, Grounding spatial named entities for information extraction and question answering, in: Kornai, A. and Sundheim, B. (eds) Proceedings of the HTL-NAACL 2003 Workshop on Analysis of Geographic References, Alberta, Canada, 2003, pp. 31–38.

[18] J. Xu, W. B. Croft, Query expansion using local and global document analysis, in: ACM SIGIR International Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 1996, pp. 4–11.

[19] C. D. Manning, P. Raghven, H. Schütze, An Intoducation To Infromation Retrieval, Cambridge University Press, Cambridge, England, 2007, draft.

[20] D. Buscaldi, P. Rosso, E. S. Arnal, Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task, in: Accessing Multilingual Information Repositories, Vol. 4022/2006, Springer (Lecture Notes in Computer Science LNCS), 2006, pp. 939–946.

[21] J. L. Leidner, Experiments with Geo-Filtering Predicates for Geographic IR, in: Accessing Multilingual Information Repositories, Vol. 4022/2006, Springer (Lecture Notes in Computer Science LNCS), 2006, pp. 987–996.

[22] R. R. Larson, F. C. Gey, V. Petras, Berkeley at GeoCLEF: Logistic Regression and Fusion for Geographic Information Retrieval, in: Accessing Multilingual Information Repositories, Vol. 4022/2006, Springer (Lecture Notes in Computer Science LNCS), 2006, pp. 963–976.

[23] T. M. Delboni, K. A. V. Dorges, A. H. F. Laender, C. A. D. Jr., Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions, in: Transactions in GIS, 2007, pp. 377–397.

[24] G. Fu, C. B. Jones, A. I. Abdelmonty, Ontology-based Spatial Query Expansion in Information Retrieval, in: On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE, Vol. 3761/2005, Springer (Lecture Notes in Computer Science LNCS), 2005, pp. 1466–1482.

[25] D. Kelly, J. Teevan, Implicit Feedback for Inferring User Preference: A Bibliography, ACM SIGIR Forum 37 (2) (2003) 18–28.

[26] T. Joachims, L. Granka, B. Pan, Accurately Interpreting Clickthrough Data as Implicit Feedback, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, 2005.

[27] D. Kelly, N. J. Belkin, Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback During Interactive Information Retrieval, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001.

[28] C. Schlieder, T. Vogele, U. Visser, Qualitative Spatial Representation for Information Retrieval by Gazetteers, in: Spatial Information Theory. Foundations of Geographic Information Science : International Conference, COSIT 2001 Morro Bay, CA, USA, September 19-23, 2001. Proceedings, Vol. 2205/2001, Springer (Lecture Notes in Computer Science LNCS), 2001, pp. 336–351.

[29] C. B. Jones, H. Alani, D. Tudhope, Geographic Information Retrieval with Ontologies of Place, in: Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science, Vol. 2205/2001, Springer (Lecture Notes in Computer Science LNCS), 2001, pp. 322–335.

[30] S. Vaid, C. B. Jones, H. Joho, M. Sanderson, Spatio-textual Indexing for Geographical Search on the Web, in: Advances in Spatial and Temporal Databases, Vol. 3633 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2005, pp. 218–235.

[31] K. Beard, V. Sharma, Multidimensional ranking for data in digital spatial libraries, International Journal on Digital Libraries 1 (1997) 153–160.

[32] D. R. F. Walker, I. A. Newman, D. J. Medyckyj-Scott, C. L. N. Ruggles, A system for identifying datasets for GIS users, International Journal of Geographical Information Systems 6 (6) (1992) 511–527.

[33] R. R. Larson, P. Frontiera, Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries, in: Research and Advanced Technology for Digital Libraries, Vol. 3232 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2004, pp. 45–56.

[34] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, B. Seefer, Design and Implementation of a Geographic Search Engine, in: Eighth International Workshop on the Web and Databases (WebDB 2005), 2005.

[35] B. Martins, M. J. Silva, L. Andrade, Indexing and Ranking in Geo-IR Systems, in: Proceedings of the ACM Workshop on Geographic Information Retrieval, ACM New York, NY, USA, 2005, pp. 31–34.

[36] M. B. Fleischman, E. Hovy, Multi-document person name resolution, in: Annual Meeting of the Association for Computational Linguistics (ACL), Reference Resolution Workshop, 2004, pp. 66–82.

[37] G. S. Mann, D. Yarowsky, Unsupervised personal name disambiguation, in: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Association for Computational Linguistics, 2003, pp. 33–40.

[38] B. E. da Graca Martins, Geographically Aware Web Text Mining, Ph.D. thesis, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa (2008).

[39] T. Rose, M. Stevenson, M. Whitehead, Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Vol. 3, 2002, pp. 827–833.

[40] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, X. Xie, GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview , in: Advances in Multilingual and Multimodel Information Retrieval, Vol. 5152/2008 of Lecture Notes in Computer Science LNCS, Springer-Verlag Berlin/Heidelberg, 2008, pp. 674–686.

[41] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, in: Journal of the American Society for Information Science, Vol. 41 of 4, John Wiley & Sons, Inc., 1990, pp. 288–297.

[42] L. Andrade, M. J. Silva, Relevance Ranking for Geographic IR, in: Proceedings of the Workshop on Geographical Information Retrieval, SIGIR'06, Seattle, USA, 2006.

[43] C. J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, 1979, 7:112-140.

[44] F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, C. Womser-Hacker, D. Santos, P. Rocha, G. M. D. Nunzio, N. Ferro, GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview, in: Evaluation of Multilingual and Multi-modal Information Retrieval, Vol. 4730 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2007, pp. 852–876.

[45] T. Mandl, F. Gey, G. D. Nunzio, N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, X. Xie, GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview , in: Working Notes for CLEF 2007, 2007.

[46] D. Buscaldi, P. Rosso, On the Relative Importance of Toponyms in GeoCLEF, in: Advances in Multilingual and Multimodel Information Retrieval, Vol. 5152/2008 of Lecture Notes in Computer Science LNCS, Springer-Verlag Berlin/Heidelberg, 2008, pp. 815–822.