Geographically constrained information retrieval
Andogah, Geoffrey

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 1

The thesis

For every house is built by someone,
but He who built all things is God.

Hebrew 3:4
(The New King James Version)

Information retrieval (IR) is the task of finding documents (usually unstructured text) that satisfy an information need from within a large collection (usually on a local computer system or the Internet) (Manning et al., 2007). Advancement in computer technology has revolutionized information generation, storage and retrieval. Millions of documents are being generated and stored daily on computers distributed across the global. Advancement in computer network technology has made it possible for millions of people to access these documents from within their personal computing systems. The Internet is the main gateway to information stored on networked computer systems across the globe. By September 30, 2009 the estimated population of the Internet users was a staggering 1.7 million people.\(^1\) The emergence of the so-called globe community is generating an enormous hunger for information about foreign places. To access information across the globe, a large percentage of the Internet users use search engines such as Google, Yahoo, Bing, Baidu, etc.

The current state-of-the-art search engines model document contents as bags-of-words consisting of words extracted from the documents. Little or no attention is paid to the document’s syntactic and semantic structure. Moreover, majority of information seekers formulate short simplistic queries consisting of two words on average (Spink et al., 2001). But there is a need

\(^1\) [05 December 2009]: http://www.internetworldstats.com/stats.htm
Chapter 1. The thesis

for the retrieval of semantically constrained information such as information about a particular locality, or in other words, for geographically constrained information retrieval.

Eighteen percent of search queries to search engines on the Internet involve some kind of geographical orientation, e.g., find documents about “oil discovery in Uganda” (Sanderson and Kohler, 2004). To satisfy these kinds of user information needs, both the non-geographical (i.e., thematic) aspect and geographical aspect of documents need consideration. This is the sphere of a novel information retrieval task known as Geographical Information Retrieval (Gey et al., 2006) which considers both the thematic and geographical aspects of the query and the documents being sought to meet the seekers’ information need. The motivating argument for Geographical Information Retrieval (GIR) is that the geographical information contained in the queries and documents may be especially important to retrieve, prioritize and visualize search results.

This introductory chapter introduces the major constituencies of a standard geographical information retrieval system before spelling out this thesis’ specific research area. Figure 1.1 shows the general processing procedure followed in geographical information retrieval task.

1.1 Geographical information retrieval

Geographical Information Retrieval (GIR) is Information Retrieval (IR) with geographical awareness added. It is poised to answers questions about the following sorts of matters (Gey et al., 2007): (1) non-geographic subjects restricted to a place (e.g., music festivals in Germany); (2) geographic subjects with non-geographic restrictions (e.g., rivers with vineyards); (3) geographic subjects restricted to a place (e.g., cities in Germany); (4) non-geographic subjects associated with a place (e.g., independence of Quebec); (5) non-geographic subjects that are a complex function of a place (e.g., European football cup matches); (6) geographical relations among places (e.g., how are the Himalayas related to Nepal? Are they inside? Do the Himalayan mountains cross Nepal’s borders?); (7) geographical relations among places associated with events (e.g., Did Waterloo occur further north than the battle of X? Were the findings of Lucy more to the south than those of the Cro-magnon in Spain?); and (8) relations between events which require a precise localization (e.g., Was it the same river that flooded last year and in which killings occurred in the XVth century?).

From these questions, we can identify four basic sorts of concepts relevant to geographically oriented questions (in addition to non-geographical
1.1. Geographical information retrieval

Figure 1.1: Common GIR processing procedure.
Chapter 1. The thesis

concepts), and these are: place-names (e.g., Germany, Quebec, Nepal, etc.), geographical relations (e.g., south, north, etc.), geographical concepts (e.g., cities, rivers, etc.) and geographical adjectives (e.g., European, etc.). These geographical terms participate in the definition of the geographical scopes of documents and queries. However, besides these there are other terms that are geographical in nature such as the names of people, the names of organisations, etc. Documents and queries which do not mention geographical terms explicitly can equally be considered to belong to a geographical scope. As a premise, it can be assumed that every document in a collection and every search query issued to a search engine has a geographical scope. Reasoning within a geographical scope domain could lead to certain geographical information retrieval processes performing better. For example, toponym resolution, query expansion and relevance ranking procedures could benefit from exploiting geographical scope information. For the work reported here geographical scope is the most important component that control the other objectives as shall be seen below. However, it should be noted that this work is not the first to consider geographical scope as an important concept to tackle geographical information retrieval problem (see also Martins (2008)).

The initial phase of a geographical information retrieval system is to mine and/or analyse geographical information concealed in documents. The next phase involves the use of the discovered information (i.e., the new knowledge) to build new models and concepts to aid in answering geographically motivated user information needs.

1.2 Research

1.2.1 Problem statement

The assumption of this thesis is that every document in a collection and every search query issued to an information retrieval system has a geographical scope. The thesis asks the question whether the geographical scopes of documents and/or search queries can be exploited to improve information retrieval results in terms of accuracy and relevance. Since the geographical scopes of documents and search queries are not explicitly specified, they must be detected for intelligent exploitation by information retrieval systems. Most of this thesis elaborates on how geographical scopes can be detected automatically and how they can be exploited once detected for toponym resolution, query expansion and relevance ranking.
1.2. Research

1.2.2 Research objective

General objective

Analyse geographical information concealed in texts for geographical information retrieval tasks.

Specific objectives

- *Mining* – Mine and/or analyse geographical information concealed in text.
- *Utilization* – Utilize mined geographical knowledge to build new models and/or concepts to satisfy geographically motivated user information needs.

The two specific objectives are further split into subordinate objectives. The *mining objective* is split into *toponym resolution* and *scope resolution* objectives. The *utilization objective* is split into *query expansion* and *relevance ranking* objectives.

The *scope resolution* objective seeks to answer the following questions:

- How can place names, geographical adjectives (i.e., adjectives referring to people and places) and names of people be used to automatically determine the geographical scope(s) of documents?  
  *Question of algorithm.*

- How well do automatically determined geographical scopes of documents compare to human assigned scopes?  
  *Question of gold standard evaluation.*

- How best can we compare the performance of scope resolution systems?  
  *Question of binary vs. non-binary evaluation metric.*

The *toponym resolution* objective seeks to address the following concerns:

- How effective is the document’s geographical scope(s) or coverage(s) in aiding the resolution of toponyms contained in the document or in other words, how effective is the document’s geographical scope in aiding the toponym resolution process?  
  *Question of heuristics.*

- How best can we compare the performance of toponym resolution systems?  
  *Question of a binary vs. a non-binary evaluation metric; and the question of reference evaluation corpora.*
Chapter 1. The thesis

Problem Statements

Figure 1.2: Schematic of the research objectives.

The query expansion objective seeks to answer the following questions:

- How effective is relevance feedback for the geographical information retrieval task?

- How effective is a scope-controlled toponym selection scheme in the relevance feedback procedure?

The relevance ranking objective seeks to answer the following question:

- How well can geographical scope and feature type information be incorporated in the document ranking procedure to prioritize documents by geography.

Figure 1.2 shows the interaction among the objectives to answer the questions raised in this thesis. The scope resolution objective is the heart of the problem-solving strategy investigated in this work. How well the scope resolution objective is answered determines how well the other objectives are answered. The scope resolution objective influences the toponym resolution objective, the query expansion objective and the relevance ranking objective.

1.2.3 Research justification

There are two main justifications for research in geographical information retrieval (GIR) – (1) infancy of GIR, and (2) user demand for geographical intelligence in conventional search engines.
Infancy of GIR

The first major campaign to seriously consider the question of geographical information retrieval was piloted by CLEF \(^2\) in 2005. Since then, three PhD dissertations have been written in the GIR domain – (1) *Toponym Resolution in Text* by Leidner (2007), (2) *Geographically Aware Web Text Mining* by Martins (2008) and (3) *Geographical Information Retrieval: Classification, Disambiguation and Modelling* by Overell (2009).

User demand

Table 1.1 shows the list of sample keywords with geographical interest of Groningen submitted to the Google \(^3\) search engine in September, 2009. Sanderson and Kohler (2004) found that 18.6% of user search queries submitted to the Excite search engine involved a geographical aspect.

The infancy of Geographically Information Retrieval (GIR), and the fact that nearly one fifth of users indicate geographical aspect in their queries is a sufficient reason for more research work in the GIR domain.

1.2.4 Contribution

This thesis’s contribution to the Geographical Information Retrieval (GIR) research is in the following areas:

1. Two new scope resolution strategies are proposed to ground a document’s geographical coverage based on the following assumptions:

\(^2\)http://clef-campaign.org/  
\(^3\)http://www.google.com
Chapter 1. The thesis

- 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).

- VIPs (i.e., political leadership) 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.

2. A new strategy that defines and model geographical scopes as a kind of document, and then uses a standard search engine library to index and search against the scope documents.

3. A new evaluation metric for the scope resolution task. The metric provides a smoother scheme than the binary scheme by weighting small discrepancies between systems. The metric incorporates the rank position information of all the correct scopes (in the result list) to evaluate systems.

4. A new toponym resolution scheme grounded on the geographical scopes assigned to documents. Other features considered in the computation are the place types (e.g., city, mountain, etc.), classification of place (e.g., administrative unit, vegetation, etc.), population size, and frequency of non-ambiguous or resolved places.

5. A new set of evaluation metrics for the toponym resolution task. The metrics incorporate the following features in their calculations: (1) the number of candidate places for a given reference, (2) the number of regions to traverse from the system resolved referent to the correct gold standard referent, and (3) the number of feature classes transversed from the system resolved referent type to the correct gold standard referent type.

6. A new scope-controlled toponym selection scheme for relevance feedback. The scheme selects toponyms to add to feedback query according
1.3. Overview

This section gives an overview of the remaining chapters in this dissertation. Chapter 2 presents an overview of the state-of-the-art on geographical information retrieval techniques in the literature. This ranges from toponym resolution, scope resolution, query expansion to relevance ranking strategies.

Chapter 3 describes the dataset and tools used to evaluate and implement various components of the Mahali system development in this work. The following datasets are described: TR-CoNLL, TR-CLEF and TR-RNW. The tools described are the named entity recognition (NER) tools. The Alias-i Lingpipe used to perform toponym recognition task is briefly described.

Chapter 4 describes the strategies proposed to handle the scope resolution problems. The strategies are grounded on two assumptions, that, (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; (2) VIPs (i.e., political leadership) in the same geographical region or at the same leadership hierarchy level tend to be mentioned together in a unit of a discourse.

Chapter 5 describes a new strategy to address the problem of toponym resolution. The toponym resolution technique described exploits the geographical scopes assigned to documents, place types (e.g., city, mountain, etc.), classification of place (e.g., administrative unit, vegetation, etc.), population size, and frequency of non-ambiguous or resolved places to accomplish the task of resolving an ambiguous place name to the location it refers to on the surface of the earth. A novel metric to evaluate the toponym resolution

7. A new relevance ranking metric that exploits a ranked list of geographical scopes assigned to query and documents. The non-geographic and geographic relevance scores are combined through a linear interpolation and, alternatively, using a weighted harmonic-mean. The harmonic mean-based combination achieved a better performance than linear interpolation. A better performance is observed when the importance attached to the non-geographical retrieval outweighs importance attached to the geographical retrieval.

All the components developed in the course of this work form the Mahali system. Mahali is a Kiswahili word meaning place.
Chapter 1. The thesis

The task is described as well. The metric is based on the argument that the number of candidate places, and closeness to the correct referent need to be factored when evaluating toponym resolution systems.

Chapters 6 and 7 describe utilization of geographical information mined in Chapters 4 and 5 for query expansion and relevance ranking. The query expansion investigates the application of relevance feedback (i.e., blind and explicit feedback) procedures to improve retrieval by adding place names found in relevant documents. The relevance ranking schemes discussed exploits the geographical scopes and feature types in documents and search queries to rank documents by geography.

Chapter 8 gives the summary of the work done in this thesis, and proposes areas that require more research in relation to what is accomplished in this thesis.