Georeferencing of Web Pages based on Context-Aware Conceptual Relationship Analysis

You-Heng Hu you-heng.hu@student.unsw.edu.au
Samsung Lim s.lim@unsw.edu.au
Chris Rizos c.rizos@unsw.edu.au
School of Surveying and Spatial Information Systems,
The University of New South Wales, Sydney NSW, Australia

Abstract. The widespread use of the Internet and the explosion of online content require the treatment of the World Wide Web (WWW) as a vast information source that satisfies the needs for many domains of interest. Recent advances in Location-Based Services, such as personal navigation and locating nearby places of interest, further facilitate the use of Web content to deliver location-sensitive information to different user groups. This trend drives the development of advanced information retrieval and search technologies that can discover, and utilise, the geographic context of Web content.

Georeferencing of a Web page is the process of assigning one or several geographic locations to that Web page. This paper addressed two of the most basic and important problems of Web page georeferencing, namely: extraction of geographic references and disambiguation of geographic references. Information extraction strategies for various internal and external geographic references are first discussed. Then a context-aware conceptual relationship analysis method is developed for geographic references disambiguation. A prototype system has been implemented and evaluated using three collections of Web pages. The impact of different types of geographic references to system performance is analysed. The results show that the proposed system works well for many well-edited Web pages. However, overall performance also depends on underlying data resources.

Keywords: Web page georeferencing, context-aware conceptual relationship analysis, geographic references disambiguation, geographic references extraction

1 Introduction

The widespread use of the Internet and the explosion of online content make it increasingly attractive to consider the World Wide Web (WWW) as a vast information source for many domains of interest. Recent advances in Location-Based Services, such as personal navigation and searching for nearby places of interest, further facilitate the use of Web content to deliver location-sensitive information.

This trend drives the development of advanced information retrieval and search technologies that can discover, and utilise, the geographic context of Web content. However, traditional search engines only build indices on the basis of the text content of Web pages. Search functionalities provided by these search engines allow users to retrieve information based on keyword matching only. Such an approach often leads to low precision and low user satisfaction (Lawrence and Giles 2000). Recently, research efforts have been devoted to the development of methodologies and technologies for information retrieval from Web pages based on their geographic context (Purves and Jones 2004). Despite differing perspectives, motivations and semantics, one of the most challenging and exciting research topics is the georeferencing of Web pages, the process of assigning one or a several geographic locations to a given Web page (McCurley 2001, Ding et al. 2000).

Many research problems must be solved to implement a robust and efficient Web page georeferencing system. In this paper two of these problems are addressed, namely the extraction of geographic references and the disambiguation of geographic references.

The first step in Web page georeferencing is to extract and map geographic references - which are the information entities that have geographic location information. Unlike plain text documents, geographic references are not only found in the content of Web pages, but also from other sources. The former are referred to as *internal* geographic references, and the latter are referred to as *external* geographic references. Some typical internal geographic references include geo-words such as place names, postal addresses, coordinates, and telephone numbers. External geographic references include the location of the host where the Web pages are stored, geographic distribution of access and the content of other related Web pages.

Recognition of geographic references is not the only challenge. Once they are extracted it is necessary to map geographic references to computer-understandable geographic objects.

However, many geographic references such as place names and location descriptions are presented using human natural language, which as a knowledge representation has a high risk of ambiguity. The procedure of mapping a geographic reference to a computer-processable object is in fact a disambiguation procedure. The *ambiguity* of geographic references has been studied widely in recent years (Smith and Crane 2001, Amitay et al. 2004, Li et al. 2003). Two types of ambiguity have been identified and analysed: geo/non-geo ambiguity and geo/geo ambiguity. Geo/non-geo ambiguity arises out the fact that a geographic reference may have some non-geographic link as well. Ambiguity between a place name and a person's name or organisation name, and ambiguity between country name and its governmental body are two examples of geo/non-geo ambiguity. Geo/geo ambiguity reflects the fact that several distinct geographic locations may have the same name. Examples of such ambiguity can be easily found all around the world. Although many algorithms and techniques have been proposed to resolve these ambiguities, their performance is severely limited due to the general lack of real-world background knowledge and a lack of the understanding of the context.

This paper presents a Web page georeferencing system that consists of several geographic references extraction strategies and a novel disambiguation method in an attempt to address the above challenges. Due to the diversity of geographic references, it is necessary to identify and develop different information extraction strategies for different types of geographic reference. These heterogeneous strategies also should be able to be integrated in a systematic and extendable fashion. In the current implementation, five types of geographic references are supported: place names, telephone numbers, postal codes, Domain Name System (DNS) records and geographic coordinates.

The geographic reference disambiguation method uses conceptual relationships derived from context-sensitive analysis. Based on a hierarchical model, three conceptual relationships are defined: identical, similar and part-of. In the disambiguation step, all possible combinations of geographic object mappings are evaluated by calculating the weighting values of the conceptual relationships found amongst them. The one that has the maximum rank value is output as the final result.

Experiments are conducted using actual Web pages to evaluate system performance. Georeferencing results are compared with those obtained by human readers. The results show that the proposed system works well for many well-edited Web pages. However, the overall performance does depend on the level of underlying data resources.

The remaining sections of this paper are organised as follows. Section 2 reviews the related research in this area. Section 3 introduces the definition of, important properties of and categorisation of geographic references. Section 4 proposes our information extraction strategies and disambiguation method. The detail implementation of our Web pages georeferencing system is presented in Section 5. Section 6 describes experiments and results. Section 7 concludes the paper and discusses the current limitations of our implementation and proposes directions for future work.

2 Related research

The GIPSY (Georeferenced Information Processing System) project (Woodruff and Plaunt 1994) was the first text document georeferencing system. The georeferencing procedure in GIPSY consisted of three steps. First, all geographic references are identified and are extracted by text matching with a thesaurus, entries of which consist of place names, coordinates and other related geographic features. Then these geographic references are mapped to geographic representations (i.e. two dimensional polygons). A weighting value which reflects its geographic nature and other properties (e.g. location and density) derived from the document content is computed and assigned to each of these polygons. The last step builds three dimensional topographies using the resulting polygons and their weighting values. The detection of the geographic scope is achieved by setting a threshold for the elevation of these topographies, and areas that are beyond some threshold are output as results. Since the inputs of GIPSY are plain text documents, only internal geographic references are extracted and used in the georeferencing process. One practical limitation of GIPSY is the use of spatial computations to construct the two dimensional and three dimensional geographic objects, which results in large memory requirements and significant processing time. Another potential problem with GIPSY is the selection of the weight matrix and threshold value, which has a direct impact on precision and recall evaluation.

The SPIRIT (Spatially-Aware Information Retrieval on the Internet) project (Jones et al. 2002) is a collaborative project with six partners across Europe. The aim of the SPIRIT project is to develop a Web-based spatially-aware search engine that allows users to specify geographic locations in their search terms. Natural language processing technologies are used widely in SPIRIT. Geographic references are first extracted from the input Web page using a NER (Named Entity Recognizer) tool that employs gazetteer-based lookup, rule-based grammar parser and machine learning approaches. SPIRIT also includes extraction rules that are derived from the mark-up format of Web pages. The extracted geographic references are mapped to geographic objects using SPIRIT ontology. Ambiguities in SPIRIT are classified into two categories: referent ambiguity and reference ambiguity. Referent ambiguity refers to the same name being potentially mapped to multiple locations. Reference ambiguity refers to a place that may have more than one name. After the disambiguation phase, one or more geographic footprints are assigned to the page. Such footprints are then encoded using XML and saved to separate files for indexing and browsing. The core element of SPIRIT is its geo-ontology. SPIRIT ontology provides semantic information that consists of geometric properties, classification categories and the relationships between geographic objects. SPIRIT also employs this geo-ontology along with other domain-specific ontologies to interpret, disambiguate and expand the geographic criteria of user queries. Unlike SPIRIT, our approach is based on a hierarchical structure that can be constructed in relatively inexpensive way.

In many previous efforts, geographic properties of places are used to disambiguate geo/geo ambiguity. In these approaches a higher weighting value is assigned to those places which are (spatially) closer to other places that have no ambiguity (Leidner et al. 2003), to those places which have a greater population (Amitay et al. 2004), and to those places which are bigger in area (Markowetz et al. 2005). These geographic properties are not included in our approach because they may not reflect the actual semantic context. Distance-based heuristics may not work when places are far away in terms of location but are close in terms of other relations. For example, even though there is a city named Sydney in Canada, when a Web page mentions "Sydney" and "Toronto", it more likely refers to two well known big cities that are located in different hemispheres. Area-based and population-based heuristics may not work when places that have a smaller area size or lower population are more important, or well known, due to any one of a number of political, economical, historical, and other, reasons.

There is a growing commercial interest in the development of Web content georeferencing technologies, which has been fuelled by high user demand of local search engines (Himmelstein 2005). Current solutions developed by major search engines such as Google Local (cf. http://local.google.com), Yahoo! Local (cf. http://local.yahoo.com) and MSN Near Me (cf. http://beta.search.msn.com) already support searching of maps and business directories based on locations. The basic user experience of these search engines are similar, users input keywords as well as a location and search radius as search criteria, and results returned by the search engine are then displayed on a map-based interface. Some search engines also extend their local search functionalities to wireless user groups. Using network-enabled mobile phones and PDAs, users can search and view Web pages, map and driving directions while moving. In contrast to our system, most of the data sources used in these search engines are obtained from existing databases, in which information on geographic locations are well organised and can be georeferenced using standard geo-coding tools and external referencing data sets such as TIGER (Topologically Integrated Geographic Encoding and Referencing system) files. However, significant and continued efforts are required to construct and maintain such databases. So far, only a few countries, such as United States, Canada and Australia, are searchable. To our knowledge none of them provides a worldwide geographically-indexed database.

3 Definition and properties of geographic references

A Web page georeferencing system takes Web pages as input, and outputs one or more geographic locations that collectively describe the geographic context of the input based on geographic references extracted from the input. In order to discuss the technical and practical aspects of such procedure, a general definition of geographic references is first given.

Definition: A geographic reference is an information entity that is discovered from the context and can be mapped to a geographic location.

A geographic reference as an information entity can be described as a tuple: *r = (g, type, value, source, index)*, in which *g* is the geographic location, *type* and *value* together describe the original information entity, *source* is the URL (Uniform Resource Locator) of the Web page, and *index* points to the position where the information entity is found. There are many way to represent *g*, a country/state/city hierarchical structure is used in this paper.

Examples of geographic references include place names, postal addresses and geographic coordinates. In general, geographic references have the following important properties:

Diversity: Unlike geographic information in many Geographic Information Systems (GIS), geographic references discovered from Web pages are not only limited to geometry and map-based representations. The Web provides multimedia communication media that support geographic information transfer over various information exchange channels such as text, digital image video and even digital audio. All of these contents can be used by humans to discover the geographic context of Web pages. In addition, geographic references also can be found from other sources related to Web pages, such as host location of the Web sites (Buyukkokten et al. 1999), geographic distribution of hyperlinks to Web pages (Ding et al. 2000), and geographic distribution of user access. However, background knowledge and the ability of heuristic reasoning are crucial for gaining a full understanding of these geographic references. Restricted by current technologies, the capability of a computer to extract and process geographic references is very limited.

Informality: Web pages are designed for humans to browse. Text written in natural language is the primary content of most Web pages. The informality of geographic references follows from the fact that natural language represents knowledge in an unstructured, but expressive manner. Relationships among information entities are also implied by the layout of content, for example, using bold typefaces or bigger font size for emphasis, or using tables and charts to relate and compare data. By contrast, machine-processable geographic information and their relationships are well defined and highly structured. Discovering formal knowledge from linguistic expressions and other resources has been an open research question for a long time. Although the emerging Semantic Web technology (Kuhn 2005) promises to enhance Web content with formal semantics, current systems and applications are limited to specific research environments only.

Redundancy: In many cases, redundant geographic references can be found within a single Web page. For instance, contact information pages of many Web sites consist of postal addresses and direction maps. Both indicate the same geographic location, and analysis of either the postal address or the direction map alone is sufficient to locate the position. Redundancies may also be found in information entities that have different geographic scale. For example, Web sites outside the United States usually have a suffix that is an abbreviation of the country name, such as .au for Australia and .cn for China. One can find country-level geographic references from the Web pages' domain name and place names mentioned in those Web pages. Some other types of information entities can provide redundancies as well, such as names of local businesses or a famous local person.

Uncertainty: The uncertainty of geographic references consists of two elements. Firstly, like much other information described using human language, geographic references have a high risk of ambiguity. Two types of ambiguities have been identified: geo/non-geo and geo/geo. Geo/non-geo ambiguity happens when a geographic reference corresponds to other non-geographic references (e.g. peoples' names, organisation names, etc). Geo/geo ambiguity reflects the fact that several distinct places share the same name. The disambiguation of geographic references is not a trivial task. However, many ambiguities can be resolved by taking advantage of context information and the redundant nature of geographic references.

The second uncertainty is due to the fuzziness of geographic concepts. To take an everyday example, how can we map the phrase "West of Sydney" to a precisely defined geometric object that can be processed by a computer? Human beings have an extraordinary ability to understand such geographic descriptions and relationships, but this ability is generally influenced by personal experiences. The answer to the above question will be different from one person to another, and at different times.

Based upon where a geographic reference appears, we propose to divide geographic references into two categories: internal geographic references and external geographic references. Internal geographic references are found within the Web page content itself. Examples of internal geographic references include textual, multimedia content and HTML metatags, which are embedded in the underlying Web page source code. Literal and sense meanings of internal geographic references are the major clues used by human readers to understand the geographic context of Web pages. In the following discussions we only focus on text-based geographic references. Other types of content, such as visual and audio information, can be accommodated in our system with adoption of advanced multimedia analysis technologies.

On the other hand, external geographic references are discovered from information sources that are outside the Web pages. Host location, geographic distribution of hyperlinks, and geographic distribution of user access are some examples of external geographic references. Normally human readers are not being able to recognise these external geographic references, because discovering such information requires particular domain knowledge and computational resources. Another type of external geographic reference is that found from related Web pages. For example, the general geographic context can be found in the "contact us" pages of company Web sites. External geographic references are useful in many situations, since they provide redundant information which is useful for disambiguation of internal geographic references.

4 Georeferencing of Web pages based on context-aware conceptual relationship analysis

4.1 Strategies for geographic references extraction

Various information extraction strategies can be applied to discover internal and external geographic references from Web pages. Strategies adopted in our system are discussed below.

Gazetteer-based text matching: A digital gazetteer is a geographic dictionary in which entries consist of information such as name, location and type designation of geographic places (Hill et al. 1999). Using a gazetteer as a filter, one can extract all words and phrases that potentially could be place names from the Web pages using text matching. This method is very simple and easy in implementation. However, it suffers from low accuracy since the semantic information within the content is not taken into account, and therefore this introduces geo/non-geo ambiguities.

Considerable efforts are required to build an extensive gazetteer. Fortunately, several public gazetteer databases are available for online access. These include the GEOnet Names Server (cf. http://www.nga.mil/gns/html/) that contains more than five million non-U.S. geographic names; the Alexandria Digital Library (ADL) Gazetteer database (Alexandria Digital Library Gazetteer 1999-) that contains about six million geographic names from around the world; and the Geographic Names Information System (cf. http://geonames.usgs.gov/index.html) that contains about two million geographic features in the United States and its territories.

Rule-based linguistics analysis: Geographic names that are not list in the gazetteer can be found using linguistics analysis based on predefined grammar rules. For example, phrases like 'the city of' and 'located at' strongly indicate that the following word or phrase is a place name, and it is not necessary to know its meaning. Other language features can be combined into this approach as well. Take English as an example, place names are mostly proper nouns, which usually begin with a capital letter. As ***Example 1*** from the introduction to Brimbank (a city close to Melbourne, Australia) shows, words like "hillside" and "sunshine" are both used as place names.

Example 1: Brimbank is a dynamic and rapidly growing city which encompasses 25 new and established suburbs including Albion, Cairnlea, Deer Park, Delahey, Hillside, Keilor, Kings Park, St Albans, Sunshine, Sydenham and Taylors Lakes….

The problem of this approach is that huge human effort is required to develop the grammar rules, and it is not suitable for multilingual applications.

Regular expression-based text matching: This approach is very useful for those geographic references that are represented in numerical format. Examples of these geographic references include telephone numbers, coordinates and postal codes. The existing global and country level numbering specifications such as North American Numbering Plan, Australia Telecommunications Numbering Plan 1997 and the United States Postal Service Zoning Improvement Plan makes it possible to

develop regular expressions to recognise them efficiently. ***Example 2*** involves a regular expression that matches the U.S. ZIP+4 Code format, which starts with five numeric digits, then a hyphen, and finishes with four numeric digits.
Example 3 illustrates a regular expression that matches the Australian telephone number format, which starts with an area code in parenthesis, followed by eight numeric digits.

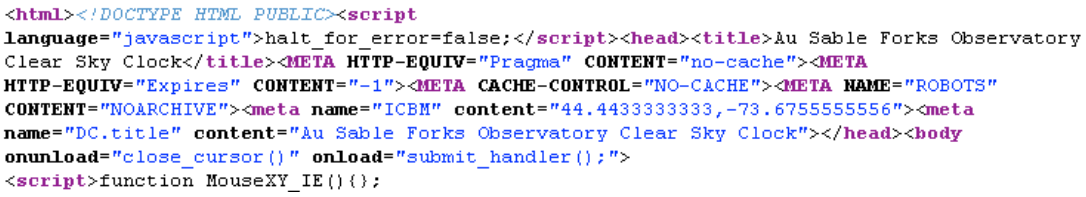
Example 2: ^[0-9]{5}-[0-9]{4}$

Example 3: ^\(0[1-9]{1}\)[0-9]{8}$

Host Location: Using the host name extracted from the URL of a Web page as a key, one can obtain the geographic location of the host from two sources. The first one is the WHOIS databases, in which host owner contact information such as postal addresses and telephone numbers can be used as geographic references. WHOIS databases are accessed using the standard WHOIS query and response protocol (Daigle 2004).

Another source is Internet Domain Name System (DNS) records. Each DNS record consists of a number of Resource Records. Two types of Resource Records can be used to determine the host location: GPOS (Farrell et al. 1994) and LOC (Davis et al. 1996). Both of these two formats use latitude/longitude/altitude coordinates to express location information. DNS records can be queried using standard nslookup tools.

Geographic MetaTag: For many HTML pages, geographic references also can be found from MetaTags embedded in the page source code. Two formats are available, ICBM addresses (cf. http://en.wikipedia.org/wiki/ICBM_address) and GeoTag addresses (cf. http://geotags.com/geo/). Locations in both formats are presented as latitude and longitude coordinates. Figure 1 illustrates an example of ICBM MetaTags, which defines a location at (44.4433333333,-73.6755555556). However, it is the authors' discretion as far as how these MetaTags are used. They can be used to refer to the location of the content, but they also can be used to refer to the location of the author(s).



```
<html><!DOCTYPE HTML PUBLIC><script
language="javascript">halt_for_error=false;</script><head><title>Au Sable Forks Observatory
Clear Sky Clock</title><META HTTP-EQUIV="Pragma" CONTENT="no-cache"><META
HTTP-EQUIV="Expires" CONTENT="-1"><META CACHE-CONTROL="NO-CACHE"><META NAME="ROBOTS"
CONTENT="NOARCHIVE"><meta name="ICBM" content="44.4433333333,-73.6755555556"><meta
name="DC.title" content="Au Sable Forks Observatory Clear Sky Clock"></head><body
onunload="close_cursor()" onload="submit_handler();">
<script>function MouseXY_IE(){};
```

Figure 1 Example of ICBM MetaTags

Using all of the above strategies, one can extract information entities that potentially are geographic references from Web pages.

4.2 Disambiguating geographic references

The disambiguating procedure in the proposed system consists of two steps. First lexical rules are adopted to remove geo/non-geo ambiguity, and then the context-aware conceptual relationship analysis method is used to further resolve remaining ambiguities.

4.2.1 Geo/non-geo ambiguity

Geo/non-geo ambiguity happens when a place name may have some non-geographic meaning also. Most of this ambiguity arises from gazetteer-based extraction, which extracts all words and phrases, regardless of their actual contextual meaning. From the perspective of semantic analysis, such ambiguity can be further subdivided into three types:

Ambiguity with other proper names: It is very common that place names are derived from the names of people and organisations. In English some surnames are taken from place names, such as York and Lancashire (Jobling 1997). On the other hand, many places in the world are named after people. Examples in Australia include Darwin, the capital city of the Northern Territory, which was named after Charles Darwin, the British naturalist (cf. http://www.darcity.nt.gov.au/aboutdarwin/history/a_brief_history.htm); and Berry, a country town of the state of New South Wales, which was named after David Berry, an early colonist (Irish 1927). One way to distinguish people's names from geographic references is to check whether the word or phrase follows a title noun (e.g. Prof, Mr, Miss, and Dr). If such title nouns exist, the extracted entity is recognised as a person's name.

Country and city names might also be included in organisations' names, to indicate their origin or headquarters. Examples include "Bank of Queensland" and the "Sydney Dance Company". Similar to people's names, one can check whether a company marker (e.g. limited, corporation, and their abbreviations) is followed by the extracted place name. If such markers exist, the entity is recognised as an organisation name.

Ambiguity with common words: Some places were named using common words. Examples in Australia include Sunshine and Waterfall. It is very difficult to disambiguate these cases using linguistic methods. One possible approach is to check whether a word begins with a capital letter. However, it is far from being reliable (Mikheev 1999). Fortunately this type of ambiguity does not occur for most capital cities and larger cities.

Used as metonymies: In many contexts place names are used to refer to another related concept, such as a governmental and community body. As ***Example 4*** from the ABC News (30th August 2004) indicates, "Australia", the country name in this context refers to its sport team.

Example 4: Australia won medals in 14 different sports in Athens with the golds spread over six sports.

Place names also can be included in attributive phrases, such as "the chef from China " and the "Prime Minister of Australia". In these examples place names are used as an attribute to their relational nouns and not necessarily to refer to any geographic location. The detection of whether a place name is used as a metonymy is the most difficult task of geo/non-geo disambiguation. Some of them can be resolved by using Part-Of-Speech (POS) tagging techniques to check whether the place name is used as a noun in the sentence.

Using the above approaches, the system determines whether an extracted entity has non-geo meaning or not. Then, entities tagged as non-geo references are eliminated from the geographic reference list. It is clear that not all geo/non-geo ambiguity can be resolved using linguistic rules-only approaches. The context-aware conceptual relationship analysis method is used in the next step to remove remaining geo/non-geo ambiguities and to resolve geo/geo ambiguities.

4.2.2 Context-aware conceptual relationship analysis

This is a well-established approach for disambiguation that employs relative information (McDonald 1996). Each information entity connects to others in its context by various relationships. Based on the country/state/city hierarchical model, three alternative relationships that are possible between two geographic objects can be found: identical, similar and part-of.

Identical: Two geographic objects are conceptually identical if they have the same country, state and city values.

It is not necessary to have two identical geographic objects using the same name. For example, both *Stalingrad* and *Volgograd* are mapped to the same famous Russia city on the west bank of Volga River. Examples also can be found from multilingual applications.

Similar: Two geographic objects are conceptually similar if: (1) they are at the same hierarchical level, and (2) they have same upper level values. For example, *Australia/New South Wales* and *Australia/Queensland* are similar, because both of them are states of Australia. *Australia* and *China* are similar, because both of them are country-level geographic objects.

Part-of: A geographic object *X* is conceptually part-of another geographic object *Y* if *X* is at lower hierarchical level of *Y*. For example, *Australia/New South Wales/Sydney* is part-of *Australia/New South Wales*, *Australia/New South Wales* is part-of *Australia*.

It is also important to note that conceptual part-of relationships are different to geographic part-of relationships. Not only geographic properties but also historical and political backgrounds must be taken into account when determining whether a conceptual part-of relationship exists between two geographic objects. One example of this issue is the state of Alaska, an exclave of the United States.

These relationships are helpful to resolve both geo/non-geo and geo/geo ambiguities. Let's consider a simple example.

Example 5: Many people think Sydney is the capital of Australia.

Three words are extracted from the above sentence as place names: *Many*, *Sydney* and *Australia*.

"*Many*" can be mapped to:

France/Lorraine/Many
United States/Louisiana/Many
…

"*Sydney*" can be mapped to many geographic objects, examples include:

Canada/ Nova Scotia/Sydney
United States/North Dakota/Sydney
United States/Florida/Sydney
Australia/New South Wales/Sydney
…

"*Australia*" can be mapped to

Australia

Adopting the context-aware conceptual relationship analysis method to the above example, firstly, "*Many*" is not a valid geographic reference since it is not related to any other entity; secondly, "*Sydney*" and "*Australia*" should be mapped to *Australia/New South Wales/Sydney* and *Australia*, because there is a part-of relationship between them and there is no relationship that can be found using other mappings.

Now, let's look at an interesting question arising from the following examples.

Example 6: China, Texas is located in Jefferson County. The latitude of China is 30.047N. The longitude is -94.335W.

Example 7: In the United States, Texas is the third largest exporter to China.

It is easy for a human reader to understand the context and then realise that the word "*China*" is mapped to two different geographic objects in the above two examples: *United States/Texas/China* for the former, and *China* for the latter. Let's use conceptual relationship analysis to find the correct one from these two possible alternative mappings.

The solution for Example 6 is quite easy. *United States/Texas/China* is part-of *United States/Texas* and that is the only relationship that can be found for the word "*China*", so "*China*" is mapped to *United States/Texas/China*.

However, for Example 7, more relationships can be found:

United States/Texas/China is part-of *United States/Texas*
China is similar to *United States*

Which mapping should be chosen? To answer this question, the concept of context distance is introduced. Given two geographic references r_1 and r_2, the context distance between r_1 and r_2 is a measure of the separation between r_1 and r_2 in the context. If r_1 and r_2 are discovered on the same Web page, the simplest way to compute context distance between them is to calculate the absolute value of the difference between $r_1.index$ and $r_2.index$.

It is found that the possibilities of $r_1.g$ and $r_2.g$ have a part-of relationship that decreases as the context distance between r_1 and r_2 increases. A place name that is used to qualify another one usually is close to the target place name, for example in Example 6 "*Texas*" is used to qualify "*China*", and in Example 7 "*United States*" is used to qualify "*Texas*". On the other hand, context distance doesn't impact on the possibility of having an *identical* or a *similar* relationship. Applying this observation, one can get the correct mapping of "*China*" in Example 7.

4.3 Algorithm outline

Given a URL (*U*) of a Web page, a Web page georeferencing system first extracts geographic references from *U*, and then disambiguates these geographic references. The results are output as a set of geographic objects.

(Step 1) Data preparation

The Web page content of *U* is first downloaded. Then a hyperlinks analyser is used to produce *RUs*, which consists of a list of URLs that are related to *U*.

(Step 2) Geographic references extraction:

(1) Use gazetteers to extract place names.

(2) Use linguistics rules to extract place names.

(3) Use regular expressions to extract telephone number, postal code and coordinates.

(4) Get the host name from *U*.

(5) Extract geographic MetaTags from the Web page header.

Geographic references are extracted from *U* and *RUs* using information extraction strategies described in Section 4.1. Extracted geographic references are input into *R*, an array of *r(g, type, value, source, index)* tuples. For each *r(g, type, value, source, index)*, *type, value, source* and *index* are initialised during the extraction procedure, *g* is initialised as NULL.

(Step 3) Use linguistic rules to eliminate of geo/non-geo ambiguity.

Each *r(g, type, value, source, index)* of *R* is checked using linguistic rules such as those discussed in the previous section if it is a place name, but it is removed if it is used in a non-geo sense. The remaining elements in *R* are passed to the subsequent mapping and disambiguation procedure.

(Step 4) Map geographic references to geographic objects.

An appropriate Web service is invoked based on the types of extracted geographic references. A geographic reference may be mapped to one or many geographic objects based on its type and value pair. These geographic objects are added to a set $G_r(g_r^1, g_r^2, ...g_r^K)$, where *K* is the total number of geographic objects in this set.

Now, the total search space *S* for disambiguation is the set of all permutations of the elements of *G*. Let *N* be the number of elements in *R*, then *S* can defined as a two-dimensional array with *N* columns and $\prod_{i=1}^{N} K_i$ rows. Each row of *S* is a possible solution.

(Step 5) Context-aware conceptual relationship analysis.

The context-aware conceptual relationship analysis method is defined as follows.

Algorithm 1 RelationshipAnalysis

```
procedure RelationshipAnalysis
input: $S$
output: solution_set

max_weight = 0.0
for each row $s(g_1^s, g_2^s, ...g_N^s)$ of $S$
  for $k$ = 1 to $N$
     $r_k.g = g_k^s$
  end for
  weight = $\sum_{i=1..N-1, j=i+1..N}$ cal_weight($r_i$ , $r_j$ )
  if (weight > max_weight) then
     clear solution_set
     add $s(g_1^s, g_2^s, ...g_N^s)$ to solution_set
     weight = max_weight
  else if weight == max_weight then
     add $s(g_1^s, g_2^s, ...g_N^s)$ to solution_set
  end if
end for
```

All possible combinations of geographic object mappings are evaluated by calculating the weighting values of the conceptual relationships found among them. After the weighting values for each row in *S* is calculated, all those $s(g_1^s, g_2^s, ...g_N^s)$ that have the maximum weighting value are placed in the *solution_set*.

The algorithm of $cal\_weight(r_i, r_j)$ is defined as follows.

Algorithm 2 *cal _ weight*

```
procedure $cal\_weight$
input: $r_i, r_j$
output: $weight$

if (( $r_i.value$ == $r_j.value$ )
    return 0.0
if ( $r_i.g$ is identical to $r_j.g$ ) then
    return $IDENTICAL\_WEIGHT$
if ( $r_i.g$ is similar with $r_j.g$ ) then
    return $SIMILAR\_WEIGHT$
if (( $r_i.g$ is part of $r_j.g$ ) or ( $r_j.g$ is part of $r_i.g$ ))
    return $PARTOF\_WEIGHT$ / (1 + lg(distance($r_i, r_j$)))
return 0.0
```

If two geographic references have the same value and type they will be assigned the same geographic object and their contribution to the solution weight is zero. Constant values *IDENTICAL_WEIGHT*, *SIMILAR_WEIGHT* and *PARTOF_WEIGHT* are determined experimentally.

The context distance between geographic references has a major impact on the part-of relationship. The distance function is defined as:

Algorithm 3 $distance(r_i, r_j)$

```
procedure $distance(r_i, r_j)$
input: $r_i, r_j$
output: $distance$

if ( $r_i.source$ != $r_j.source$ ) then
    return $EXTERNAL\_DISTANCE$
else
    if (( $r_i.type == DNS$ ) or ( $r_j.type == DNS$ )) then
        return $DNS\_DISTANCE$
    end if
    if (( $r_i.type == METATAG$ ) or ( $r_j.type == METATAG$ )) then
        return $METATAG\_DISTANCE$
    end if
    return $\sqrt{(src\_distance(r_i, r_j))^2 + (L \times level\_distance(r_i, r_j))^2}$
end if
```

If the two geographic references are extracted from different sources, the distance between them is a constant value *EXTERNAL_DISTANCE*, otherwise, constant values *DNS_DISTANCE* and *METATAG_DISTANCE* are returned if one of these two geographic references is a DNS or a MetaTag. In the case when the two geographic references are both extracted from the Web page body content, the distance between them is defined as a function of *src_distance* and *level_distance* , in which *src_distance* is the difference in their *index* value, and the $L \times level\_distance$ is the difference between their positions in the HTML tree structure.

(Step 6) Validate results.
Two types of geographic references *r* are tagged as invalid: (1) those that have two or more different mapping in *solution_set;* and (2) those have no relationship with others. Remaining entries of *solution_set* are output as the final results.

5 Implementation

5.1 System overview

The architecture of our Web page georeferencing system is shown in Figure 2.The input of the system is the URL of a Web page, and the output is a collection of geographic objects described in a country/state/city hierarchical structure. The system was implemented using the JAVA language.

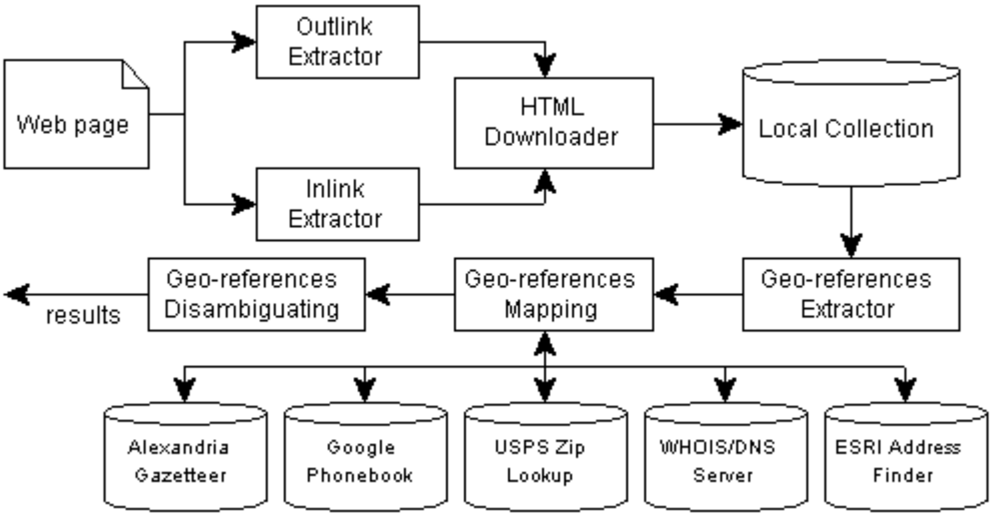


Figure 2 Web page georeferencing system architecture

5.2 Building a local Web page collection

Building a local Web page collection is the first step of georeferencing. The local collection consists of the Web page of the input URL, which we refer to as the *original page,* and a selection of Web pages that link to or from the input URL, which we refer to as *external pages*.

An original page is first downloaded and saved to a local file system. Next, a HTML parser is used to extract all hyperlink elements from the original page, and place them in a temporary Web link array *out_links*.

To get Web pages that link to the input URL, the "Find web pages that link to" feature of the Google search engine is used. Query results returned from Google are placed in another temporary hyperlink array *in_links*.

out_links and *in_links* are then merged together after duplicated entries are removed. Hyperlinks that link to multimedia content such as images, sound and video files are removed as well. The remaining entries are placed in a hyperlink array *external_links*.

Finally, a number of Web links are randomly selected from *external_links*. The selected Web links are downloaded and saved to a local file system as *external pages*.

5.3 Extracting geographic references

The second step is to extract geographic references from the local Web page collection. This functionality is implemented by extending the GATE system developed at the Natural Language Processing Research Group at the University of Sheffield (Cunningham 2002). The GATE system provides a general architecture for text engineering related applications. GATE consists of a collection of built-in process resources (e.g. tokeniser, sentence splitter, POS tagger and Named Entity Recogniser) and language resources (e.g. gazetteer and grammar rules). GATE is invoked through its application programming interface.

GATE is used to extract location names, postal codes and telephone numbers. To extract place names, GATE supports the gazetteer-based lookup method and grammar rules-based geo/non-geo disambiguation. Regular expressions in GATE are written in the JAPE (Java Annotation Patterns Engine) language.

The GATE built-in gazetteer includes names of countries, capital and major cities around the world, and many places in Europe. The built-in JAPE grammar rules for postal codes and telephone numbers are based on the England style. In the current implementation GATE is extended with more a detailed worldwide gazetteer, and regular expressions for U.S style telephone numbers and ZIP codes.

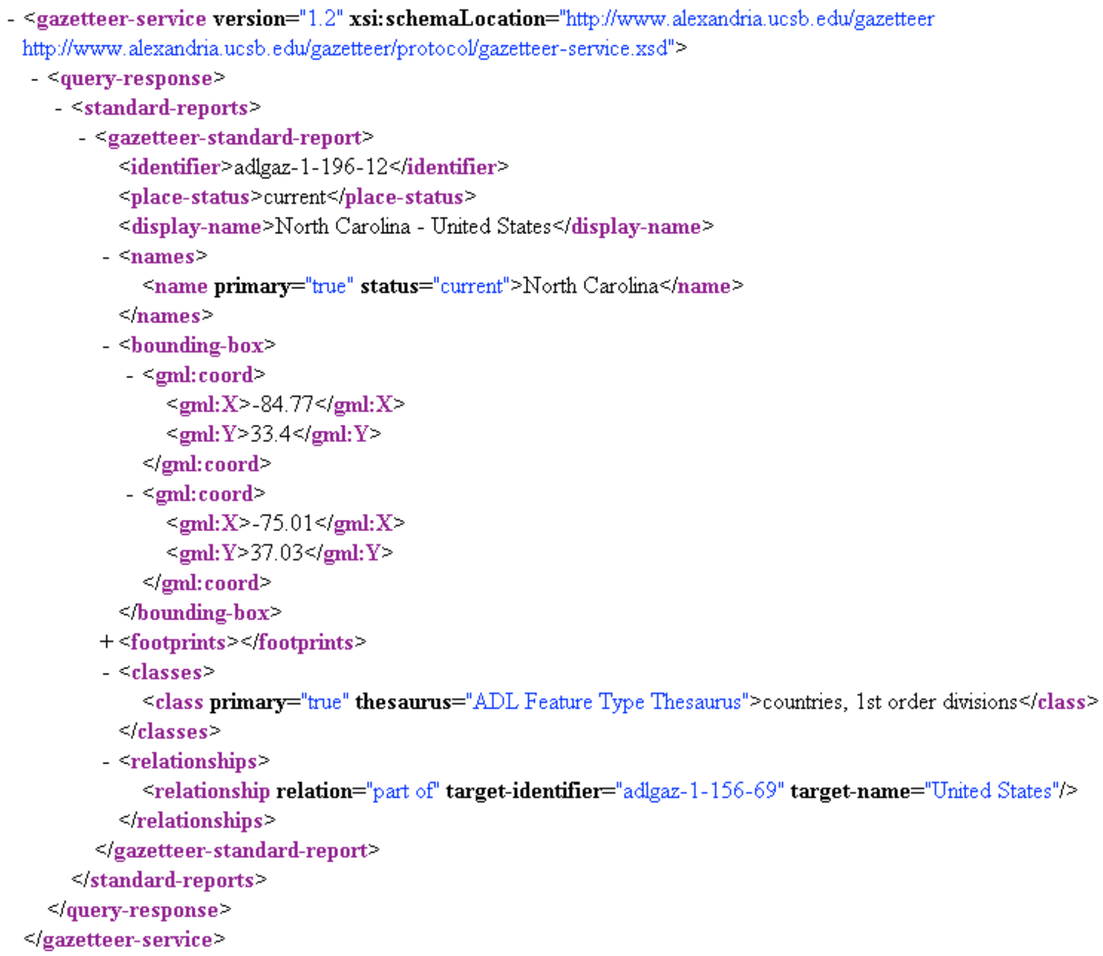
GATE only provides the position of the extracted entities in the Web page source code. A HTML parser is used to retrieve their position in the HTML tree structure. The HTML parser is also used to extract geographic MetaTag elements from the Web pages. Two types of geographic MetaTags are supported: ICBM addresses and GeoTags.

5.4 Mapping geographic references

The third step is to map the extracted geographic references to the country/state/city hierarchical structure. The current implementation supports mapping of five types of geographic references:

(1) Place names:

The Alexandria Digital Library Gazetteer is used to map place names. The ADL Gazetteer database can be accessed by client applications through the ADL Gazetteer Protocol (Smith 1996). To map a location name, the system sends a <*name-query*> message to the ADL Gazetteer server, then retrieves the qualified <*display-name*> element from the response message and splits it into an array of strings, which is used to map the place name to a country/state/city structure. This array is called the *name_array*. The <*classes*> elements of the response message are used to filter non-related geographic objects, such as agricultural and industrial sites. Figure 3 shows an example of the response message for the place name "*North Carolina*".

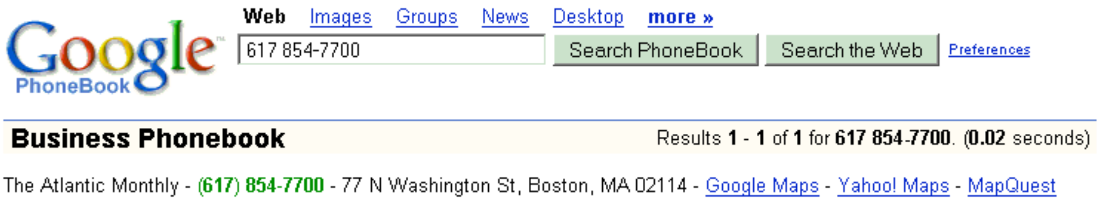


```
- <gazetteer-service version="1.2" xsi:schemaLocation="http://www.alexandria.ucsb.edu/gazetteer
  http://www.alexandria.ucsb.edu/gazetteer/protocol/gazetteer-service.xsd">
  - <query-response>
    - <standard-reports>
      - <gazetteer-standard-report>
          <identifier>adlgaz-1-196-12</identifier>
          <place-status>current</place-status>
          <display-name>North Carolina - United States</display-name>
        - <names>
            <name primary="true" status="current">North Carolina</name>
          </names>
        - <bounding-box>
          - <gml:coord>
              <gml:X>-84.77</gml:X>
              <gml:Y>33.4</gml:Y>
            </gml:coord>
          - <gml:coord>
              <gml:X>-75.01</gml:X>
              <gml:Y>37.03</gml:Y>
            </gml:coord>
          </bounding-box>
        + <footprints></footprints>
        - <classes>
            <class primary="true" thesaurus="ADL Feature Type Thesaurus">countries, 1st order divisions</class>
          </classes>
        - <relationships>
            <relationship relation="part of" target-identifier="adlgaz-1-156-69" target-name="United States"/>
          </relationships>
        </gazetteer-standard-report>
      </standard-reports>
    </query-response>
  </gazetteer-service>
```

Figure 3 ADL Gazetteer response message

(2) Telephone numbers:

The Google reverse phone number lookup feature is used to achieve the task of mapping telephone numbers. Figure 4 shows an example of such a query. Due to the limitation of the Google database, only U.S. telephone numbers are supported.



Business Phonebook

Results **1** - **1** of **1** for **617 854-7700**. (**0.02** seconds)

The Atlantic Monthly - (617) 854-7700 - 77 N Washington St, Boston, MA 02114 - Google Maps - Yahoo! Maps - MapQuest

Figure 4 Google phone number lookup

(3) Postal codes:

The Address Information APIs provided by the USPS (The United States Postal Service) is used to map postal codes. It is part of the USPS Web Tools that support the lookup of city and state information using ZIP codes for the U.S.

(4) Host locations:

Two information sources are used to map domain names. The first one uses contact information in the WHOIS record. The second one uses location information in the DNS record.

(5) Geographic coordinates:

Mapping a geographic coordinate to an address is called reserve geo-coding. The Address Finder service provided by ESRI (Environmental Systems Research Institute, Inc.) is used to obtain addresses based on longitude and latitude coordinates. Locations in the U.S., Canada, Australia and New Zealand are supported.

All of the above methods are integrated into the system through service provider APIs or screen scraper technologies. It is can be easily extended to add more online information services to support other countries and other types of geographic references. The extracted geographic references and their mappings are then fed to the disambiguation module.

5.5 Disambiguating

As discussed above, the total computation cost is dependent on the search space, hence a large number of geographic references and their potential mappings may increase computational time and memory requirements dramatically. For practical reasons, any geographic reference that is a place name and its *name_array* consists of more than three elements is eliminated in order to reduce the search space.

Taking the extracted geographic references and their mappings as input, the disambiguating algorithm grades each of the possible combinations of geographic reference candidates. Two geographic objects are matched by comparing their *name_array* in a case-insensitive fashion. The results are output as a collection of key/value pairs in which the key is the text string of a geographic reference and the value is a country/state/town hierarchical structure.

6 Evaluation

This section describes the experiments undertaken and presents some results. Evaluation metrics for the experiments are based on precision and recall measurements. The testing corpus consists of three collections of actual Web pages.

6.1 Evaluation metrics

Two well understood measures are adopted for our experiments, precision (*P*) and recall (*R*), where:

$$P = \frac{\text{number of correct results}}{\text{number of results}} \text{ and } R = \frac{\text{number of correct results}}{\text{number of proved answers}}$$

Precision measures the number of correct results compared to the number of results, whereas recall measures the number of correct results compared to the number of proven answers.

Proven answers of a Web page were provider by a human reader. The reader studied the Web page and assigned one or more geographic objects that reflect the Web page geographic context. These geographic objects were recorded in the country/state/city hierarchical model as proven answers.

The output of the Web page georeferencing system is a set of geographic references and their geographic object mappings. The outputs are compared with the proven results, a correct result is one where all values at country, state and city levels are correct.

6.2 Testing corpus

The testing corpus consists of three collections of Web pages, each of them consisting of 50 original *pages* and each *original page* has 10 *external pages*. The total number of Web pages that were processed was 1309, and the total file size was 52.6M bytes.

The first collection is referred to here as "News Collection", consisting of Web pages taken at random from CNN online world news (January 2004 to April 2004). These Web pages are written in a good grammar and sentence structure style. Geographic contexts of these Web pages are obvious and can be found easily from the first few sentences.

The second collection is referred to as the "GeoURL Collection", consisting of Web pages taken at random from the GeoURL ICBM Address Server (http://geourl.org). All of these Web pages use ICBM MetaTags. Most of them are Weblog pages and are written in a relatively casual style.

The last collection is referred to here as the "Arbitrary Collection". We used the method described in (Amitay et al. 2004) to generate this collection. Web pages are randomly selected from Google query results of "+the", "+and" and "+in".

During the generation of the testing corpus, only HTML Web pages are selected, other types of links, such as XML, PDF and Word documents are ignored. We also eliminated those:

- Content size less than 3K bytes.
- Not written in English.
- Using HTML frames.

6.3 Experimental results

Geographic reference extraction strategies are applied to the testing corpus first. As depicted in Table 1, the average number of extracted geographic references per Web page ranged from 11.95 to 19.48. MetaTags were found from the GeoURL collection only. Phone numbers and postal codes were found from the GeoURL and Arbitrary collections, with most of them being part of the contact information of companies and organisations.

As discussed above, some of these geographic references, in particular place names, have ambiguities. Table 2 shows the statistical data for number of distinct place names and their gazetteer entries. The average number of gazetteer entries per place name was similar for all three collections, which is about 4 entries per place name.

Another interesting finding was that while the News collection contains the highest number of extracted place names, its distinct place names number was the lowest one among all thee collections. The reason behind this is that for a period of time, most of the world news stories focused on a limited number of “hot” places, such as "Gaza", "Iraq" and "Kosovo" in our case.

Table 1 Extracted geographic references

	News	GeoURL	Arbitrary	Total
Number of Web Pages	521	334	454	1309
Place Name	9629	5668	4837	20134
MetaTag	0	200	0	200
Phone No.	1	143	128	272
Postal code	0	0	5	5
DNS	521	334	454	1309
Total number	10151	6345	5of 424	21920
Geographic references per page	19.48	19.00	11.95	16.75

Table 2 Place names and their Gazetteer entries

	News	GeoURL	Arbitrary	Total
Distinct Place Name	418	1223	1484	3125
Gazetteer entry	1846	4664	5748	12258
Entries per name	4.42	3.81	3.87	3.92

We applied context-aware conceptual relationship analysis with the following parameters:

- A weighting value of *identical* relationship *IDENTICAL_WEIGHT* = 4.0;
- A weighting value of *similar* relationship *SIMILAR_WEIGHT* = 4.0;
- A weighting value of *part-of* relationship *PARTOF_WEIGHT* = 2.0;
- A distance value of external geographic references *EXTERNAL_DISTANCE* = 500;
- A distance value of DNS geographic references *DNS_DISTANCE*= 500;
- A distance value of MetaTag geographic references *METATAG_DISTANCE* = 500; and
- A distance value of HTML tree nodes *L* = 10.

Table 3 Performance of context-aware conceptual relationship analysis

	News		GeoURL		Arbitrary	
	P(%)	R (%)	P (%)	R (%)	P (%)	R (%)
Internal Evidences	73.18	75.20	25.33	46.34	36.71	49.15
Internal Evidences DNS MetaTag	73.18	75.20	22.83	51.22	34.83	52.54
Internal Evidences DNS MeaTag External Evidences	73.56	75.59	22.35	46.34	34.83	52.54

The precision and recall measures of each collection are summarised in Table 3. For comparison, the impact of different types of geographic references is determined. Firstly, only internal geographic references were used in the disambiguating procedure. Relatively high precision and recall values were found in the News collection (P= 73.18%, R = 75.20%). Significantly lower results were observed in the GeoURL (P= 25.33%, R = 46.34) and Arbitrary (36.71%, 49.15%) collections.

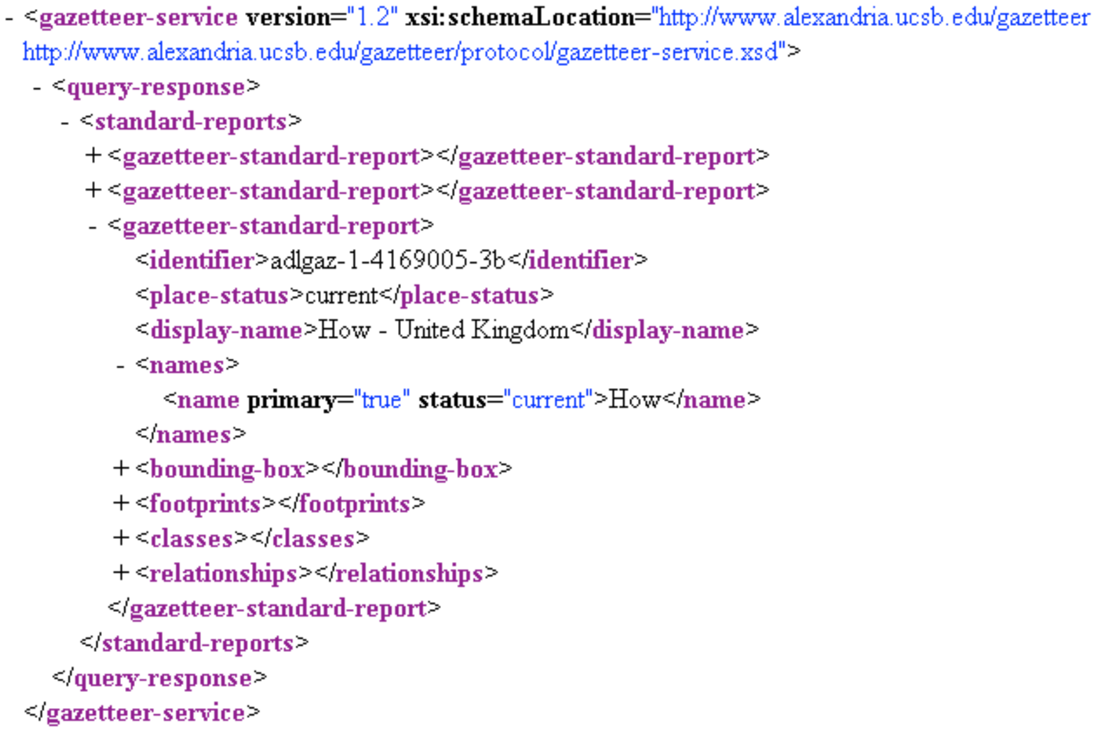
Next DNS and MetaTag are added. The addition of these data had no impact on the News collection. For the GeoURL and Arbitrary collections however, recall increased 10.53% and 6.90% respectively; but their precision decreased 9.87% and 5.12% respectively.

Finally external geographic references are added. The addition of these data slightly improved the precision (by 0.52%) and recall (0.52%) of the News collection. For the GeoURL collection precision decreased again, by 2.10%, and recall decreased by 9.53%. Precision and recall didn't change for the Arbitrary collection.

Several observations are noteworthy from these experiments. Firstly, the ambiguity of place names has global impact on various types of Web pages. For Web pages of the News collection, which are well edited and well organised, internal evidence extracted from the Web page content are the primary cues to disambiguate. Location of the news Web server is not relevant to news stories in many cases, with no impact on the system performance. On the other hand, DNS and geographic MetaTags data are very useful for the GeoURL and Arbitrary collections, although they also introduce some error results.

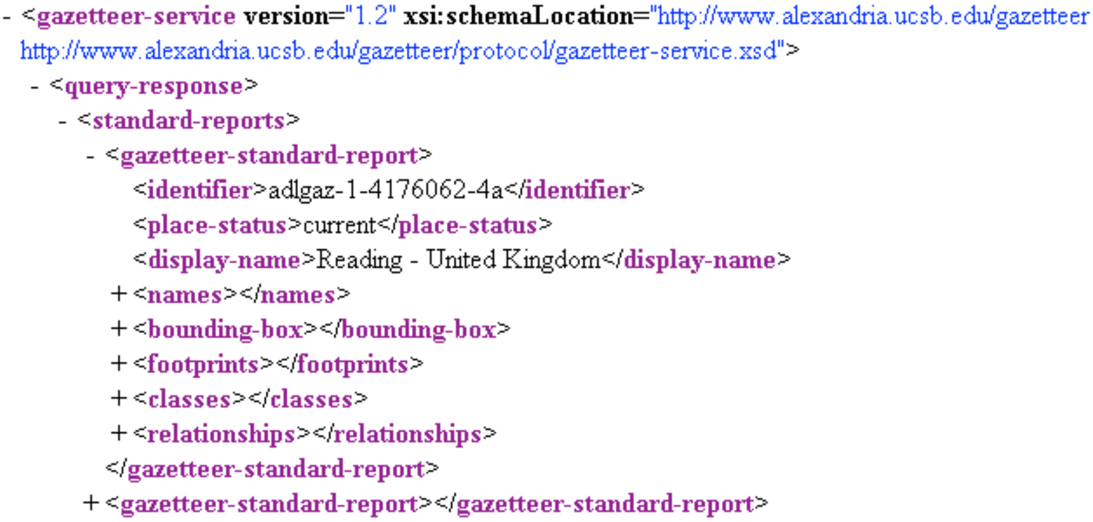
Secondly, the effects of those geographic references extracted from external Web pages are not as we thought. Most Web pages in the News collection are linked to and from other Web pages that have related content; therefore external geographic references can be used to disambiguate internal geographic references. However, for the GeoURL collection, links may lead to external Web pages that have totally different location orientation, which caused the decrease of both precision and recall. The negative effect induced by external geographic references shows that we should not add them to the search space.

Thirdly, it is not completely satisfactory to determine relationships between geographic references by using simple string matching of their qualified names. The semantic nature of places' qualified names is not well defined in the Alexandria Gazetteer, and some of the qualified names do not follow the city/state/country structure. For example, the following XML documents (see Figure 5 and Figure 6) are gazetteer entries of “*How*” and “*Reading*”, both of these two words can be mapped to state-level geographic objects of the United Kingdom. Such misleading information makes the data noisier, and therefore impacts on the performance of the algorithm.



```
- <gazetteer-service version="1.2" xsi:schemaLocation="http://www.alexandria.ucsb.edu/gazetteer
  http://www.alexandria.ucsb.edu/gazetteer/protocol/gazetteer-service.xsd">
  - <query-response>
    - <standard-reports>
      + <gazetteer-standard-report></gazetteer-standard-report>
      + <gazetteer-standard-report></gazetteer-standard-report>
      - <gazetteer-standard-report>
          <identifier>adlgaz-1-4169005-3b</identifier>
          <place-status>current</place-status>
          <display-name>How - United Kingdom</display-name>
        - <names>
            <name primary="true" status="current">How</name>
          </names>
        + <bounding-box></bounding-box>
        + <footprints></footprints>
        + <classes></classes>
        + <relationships></relationships>
        </gazetteer-standard-report>
      </standard-reports>
    </query-response>
  </gazetteer-service>
```

Figure 5 Alexandria Gazetteer entry of “How”



```
- <gazetteer-service version="1.2" xsi:schemaLocation="http://www.alexandria.ucsb.edu/gazetteer
  http://www.alexandria.ucsb.edu/gazetteer/protocol/gazetteer-service.xsd">
  - <query-response>
    - <standard-reports>
      - <gazetteer-standard-report>
          <identifier>adlgaz-1-4176062-4a</identifier>
          <place-status>current</place-status>
          <display-name>Reading - United Kingdom</display-name>
        + <names></names>
        + <bounding-box></bounding-box>
        + <footprints></footprints>
        + <classes></classes>
        + <relationships></relationships>
        </gazetteer-standard-report>
      + <gazetteer-standard-report></gazetteer-standard-report>
```

Figure 6 Alexandria Gazetteer entry of “Reading”

Finally, our proposed method requires considerable computational time. The complexity of a complete search is $O(N^G)$, where N is the number of geographic references found from content, and G is the number of possible maps of each geographic reference. Given the statistical data in Table 1 and Table 2, the average value of N is 17 and the average value of G is 4. Taking into account other computational tasks, such as downloading of Web pages, invoking of GATE to extract geographic references and invoking of Web services to map geographic references, it is apparent that our approach is only appropriate for post-processing applications.

Taken together these observations suggest that although the conceptual relationships are very useful for Web pages georeferencing, the overall performance depends significantly on the underlying data.

7 Conclusion and future work

This paper proposed a Web page georeferencing system that addressed two basic and important problems of Web page georeferencing: extraction of geographic references and disambiguation of geographic references. In the proposed system, internal and external geographic references are extracted and mapped first using various information extraction strategies, and then a context-aware conceptual relationship analysis method is developed for geographic reference disambiguation. Three conceptual relationships between geographic references are defined and used in the disambiguation method. A prototype has been implemented and evaluated on three collections of actual Web pages. The results show that the system achieved high precision and recall for many well-edited Web pages. The impact of different types of geographic references on system performances was analysed. The importance of the underlying data resources on system performance was highlighted.

Future work is planned in the following:

- Reduce the search space of disambiguation. As discussed above, the major drawback of the proposed method is its high computational requirements. However, we believe that the search space can be limited by applying a window size, and only geographic references in the window need be involved in the disambiguation procedure.

- After all geographic references are extracted and disambiguated, it is necessary to rank them. Consequently, future work should include the development of spatial ranking methods for Web pages, in which not only the position and occurrence frequencies, but also other HTML features such as font (eg. style and size), positions in HTML tree structures, etc., should be taken into account.

- We are also planning to combine our this georeferencing method with traditional information retrieval technologies to develop a Geographic Information Retrieval (GIR) system, and evaluating its performance using a comparable method, such as the GeoCLEF track of CLEF 2005 (Cross Language Evaluation Forum), which uses ad hoc information retrieval tasks in standard formats, and is based on large scale of document collections.

References

Alexandria Digital Library Gazetteer (1999-) Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents. http://www.alexandria.ucsb.edu/gazetteer (visited 10th September, 2005).

Amitay E, Har'El N, Sivan R and Soffer A (2004) In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Sheffield, United Kingdom, pp. 273-280

Buyukkokten O, Cho J, Garcia-Molina H, Gravano L and Shivakumar N (1999) In Proceedings of Second International Workshop on the Web and Databases (WebDB). Philadelphia, Pennsylvania, USA, pp. 91-96.

Cunningham H (2002) GATE, a General Architecture for Text Engineering. Computers and the Humanities, 36: 223-254.

Daigle L (2004), RFC 3912 - WHOIS Protocol Specification. http://www.faqs.org/rfcs/rfc3912.html (visited 10th September, 2005).

Davis C, Vixie P, Goodwin T and Dickinson I (1996), RFC 1876 - A Means for Expressing Location Information in the Domain Name System. http://www.faqs.org/rfcs/rfc1876.html (visited 10th September, 2005).

Ding J, Gravano L and Shivakumar N (2000) In Proceedings of the 26th International Conference on Very Large Data(VLDB). BasesMorgan Kaufmann Publishers Inc., Cairo, Egypt, pp. 545-556.

Farrell C, Schulze M, Pleitner S and Baldoni D (1994),RFC 1712 - DNS Encoding of Geographical Location. http://www.faqs.org/rfcs/rfc1712.html (visited 10th September, 2005).

Hill L L, Frew J and Zheng Q (1999),Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. http://www.dlib.org/dlib/january99/hill/01hill.html (visited 2005).

Himmelstein M (2005) Local Search: The Internet Is the Yellow Pages. Computer, 38: 26-34.

Irish C A (1927) Names of Railway Stations in New South Wales. With their Meaning and Origin. Royal Australian Historical Society, 13: 99-144.

Jobling M A (2001) In the name of the father: surnames and genetics. Trends Genet, 17: 353-357.

Jones C B, Purves R, Ruas A, Sanderson M, Sester M, Kreveld M v and Weibel R (2002) In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Tampere, Finland, pp. 387-388

Kuhn W (2005) Geospatial Semantics: Why, of What, and How? Lecture Notes in Computer Science, 3534: 1-24.

Lawrence S and Giles C L (2000) Accessibility of information on the Web. Intelligence, 11: 32-39.

Leidner J L, Sinclair G and Webber B (2003) Grounding spatial named entities for information extraction and question answering. In: HLT-NAACL 2003 Workshop on the Analysis of Geographic References. Online Proceedings, available from http://gunsight.metacarta.com/kornai/NAACL/WS9/Conf/.

Li H, Srihari R K, Niu C and Li W (2003) InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: HLT-NAACL 2003 Workshop on the Analysis of Geographic References. Online Proceedings, available from http://gunsight.metacarta.com/kornai/NAACL/WS9/Conf/.

Markowetz A, Chen Y-Y, Suel T, Long X and Seeger B (2005) In Proceedings of the 8th International Workshop on the Web and Databases(WebDB). Baltimore, Maryland, USA, pp. 19-24.

McCurley K S (2001) In Proceedings of the 10th international conference on World Wide Web (WWW).ACM Press, Hong Kong, China, pp. 221-229.

McDonald D D (1996) In Corpus processing for lexical acquisition(Eds, Boguraev, B. and Pustejovsky, J.) MIT Press, pp. 21-39.

Mikheev A (1999) A knowledge-free method for capitalized word disambiguation. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 159-166.

Purves R and Jones C (2004) Workshop on Geographic Information Retrieval, SIGIR 2004. http://www.sigir.org/forum/2004D/purves_sigirforum_2004d.pdf (visited 10th September, 2005).

Smith D A and Crane G (2001) Disambiguating Geographic Names in a Historical Digital Library. In Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, pp. 127-136.

Smith T R (1996) A Digital Library for Geographically Referenced Materials. Computer, 29: 54-60.

Woodruff A G and Plaunt C (1994) GIPSY: Georeferenced Information Processing System. Journal of the American Society for Information Science, 45: 645-655.