Geoinformatica (2011) 15:609–631
DOI 10.1007/s10707-010-0118-z

Ontology-driven discovery of geospatial evidence in web pages

Karla A. V. Borges · Clodoveu A. Davis Jr · Alberto H. F. Laender · Claudia Bauzer Medeiros

Received: 22 December 2009 / Revised: 26 August 2010
Accepted: 19 October 2010 / Published online: 3 November 2010

Abstract When users need to find something on the Web that is related to a place, chances are place names will be submitted along with some other keywords to a search engine. However, automatic recognition of geographic characteristics embedded in Web documents, which would allow for a better connection between documents and places, remains a difficult task. We propose an ontology-driven approach to facilitate the process of recognizing, extracting, and geocoding partial or complete references to places embedded in text. Our approach combines an extraction ontology with urban gazetteers and geocoding techniques. This ontology, called OnLocus, is used to guide the discovery of geospatial evidence from the contents of Web pages. We show that addresses and positioning expressions, along with fragments such as postal codes or telephone area codes, provide satisfactory support for local search applications, since they are able to determine approximations to the physical location of services and activities named within Web pages. Our experiments show the feasibility of performing automated address extraction and geocoding to identify locations associated to Web pages. Combining location identifiers with basic addresses improved the precision of extractions and reduced the number of false positive results.

K. A. V. Borges
PRODABEL-Empresa de Informática e Informação do Município de Belo Horizonte,
Av. Pres. Carlos Luz, 1275, 31230-000 Belo Horizonte, MG, Brazil
e-mail: karla@pbh.gov.br

K. A. V. Borges · C. A. Davis Jr (✉) · A. H. F. Laender
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais,
Av. Pres. Antônio Carlos, 6627, 31270-010 Belo Horizonte, MG, Brazil
e-mail: clodoveu@dcc.ufmg.br

A. H. F. Laender
e-mail: laender@dcc.ufmg.br

C. B. Medeiros
Instituto de Informática, Universidade de Campinas,
Av. Albert Einstein,1251, 13083-970 Campinas, SP, Brazil
e-mail: cmbm@ic.unicamp.br

Keywords Geographic information retrieval · Extraction ontologies · Geospatial evidence in text · Positioning expressions · Geocoding

1 Introduction

Web pages often contain geospatial evidence such as place names, addresses, postal codes, and phone numbers, in a semi-structured fashion. Humans are able to recognize and use such information and attribute geographic meaning to Web pages, as part of the tedious and time-consuming process of filtering search engine results to fulfill their needs for information. On the other hand, automatic recognition of geographic characteristics embedded in Web data and documents still remains a difficult task.

People are always looking for Web pages containing useful information about everyday tasks. Local merchants, services, and news are frequently sought [24]. The Web has the potential to provide more efficient local information access for consumers ("find me a nearby restaurant") and tourists ("show me attractions located within 1,000 meters"), but current search tools underuse its potential as a repository of geographic information [7, 27, 37]. According to Schockaert et al. [37], current local search services, such as Google Maps[1] or Yahoo! local,[2] process queries against a fixed and structured list of businesses. In spite of much evolution in the last few years, such services still are not able to perform geographic searching over the unstructured or semi-structured contents that are usual on the Web, thus people use search engines as an alternative. However, a sentence conveying geographic intent submitted as a set of keywords still leads to misinterpretation. For instance, if a user inputs "hotel far from downtown New York", results are likely to include New York hotels referenced in sites which contain the words "downtown" and "far", but the actual location of the hotels varies. Mapping, routing capabilities, and functions to locate businesses on maps or on high-resolution satellite imagery make visualization and interaction easier [32], but do not actually solve search limitations as to user-intended geographic scope.

Nevertheless, when the user needs to find something on the Web that is related to a place, chances are place names will be submitted along with some other keywords to a search engine. Sanderson and Kohler [36] verified that about 18% of keywords submitted as queries to the Excite search engine contained geography-related terms. In Brazil, a six month analysis of query logs from TodoBR (a major Brazilian search engine, acquired by Google in 2005) [13] revealed that 14.1% of queries contained at least one geography-related term, such as the name or type of a place, a spatial relation, or a word indicating locality. The study also showed that at least 20% of the Web pages contained one or more easily recognizable, unambiguous geographic identifiers, such as postal addresses, and included locally relevant content.

The recent availability and growing popularity of Web-based mapping services in cellular phones is an example that shows the usefulness of better search tools for a local scope [44]. Sites such as YouTube increasingly use global coordinates and place names as part of the content selection resources, improving search by combining location references to the well-known keyword-based approach employed by search engines. Recognizing local geographic references embedded in Web data sources is also important for applications that support online social interaction, such as blogs (Blogger, Windows Live

[1] http://maps.google.com
[2] http://local.yahoo.com

Spaces, Twitter), friend finders (Whereyougonnabe, Reunion, Classmates), interest groups (Buzznet, Flixster), content sharing (Flickr, YouTube), friendship networking (Orkut, Facebook, MySpace, hi5), and many others. Increased understanding of reference context and semantics is necessary to fulfill this growing demand [16].

This article presents an ontology-based approach that aims to recognize, extract, and geocode geospatial evidence with local (urban) characteristics, such as street names, urban landmarks, telephone area codes, and postal addresses. It focuses on extracting geographic knowledge from local Web business or service pages. Our objective is to provide support to location-based services integrating Web pages with urban locations. This meets the users' growing demand for such services, and has vast commercial, economic, and social applications

The remainder of this article is organized as follows. Section 2 presents related work. Section 3 discusses how geospatial evidence with local characteristics can be recognized and extracted from Web pages. Section 4 presents OnLocus, an ontology developed to extract geospatial evidence from Web pages. Section 5 shows results obtained from experiments using OnLocus on Brazilian Web pages. Section 6 presents conclusions and directions for future work.

2 Related work

Geographic aspects of the Web can be explored using two approaches [2]. The first approach, *Source Geography,* uses Internet infrastructure elements to obtain information about the physical location of hosts, which are then used as a rough approximation for the user's location. This approach can lead to imprecise and incorrect locations, since the user can be connected to a remote server, and pages referring to a location can be stored in servers located elsewhere. The Source Geography approach can also be applied in the case of mobile users, in which the location approximation is obtained using various signal processing and network-based techniques [45]. Using locations based on IP addresses can be imprecise, but techniques used in mobile computing, which use GPS devices embedded in smartphones or WiFi-based triangulation can be much more accurate.

The second approach, *Target Geography*, uses elements contained in the page to deduce locations. Such elements include place names, postal addresses, and phone numbers. The challenge for Target Geography involves evidence extraction, semantic analysis, and interpretation, in order to link Web pages to geographic locations. Previous works [2, 20, 25, 31, 40, 48, 50] have considered using the intended meaning of terms, expressions, and phrases in natural language as a useful paradigm for navigating and retrieving Web geographic information [14].

Geographic Information Retrieval (GIR) is an applied research field that involves indexing, searching, retrieving, and browsing georeferenced information sources, and designing systems to execute these tasks effectively and efficiently [28]. As compared with Information Retrieval (IR), GIR assumes that some geography-related semantic information is present, in the form of geographic metadata or by the incorporation of semantic notions about spatial relationships and location. Like IR, GIR includes indexing, storage and ranking, but browsing requires more sophisticated interfaces (usually Web maps). The recognition and manipulation of place names, which can function as a special set of keywords, is also critical for GIR. For that purpose, *gazetteers* (dictionaries of place names) are often used [23, 42].

Fu et al. [20] and Silva et al. [40] use geographic ontologies to obtain spatial metadata from Web pages. Based on knowledge from the ontologies and gazetteers, seen here as

location knowledge bases, each geographic term found in the page is extracted and linked to a *spatial footprint*. Footprints associated with the page are then used to build a spatial index for a search engine. An *extraction ontology* [17] is defined as an instance of a conceptual model which describes an application in a given domain of interest in terms of a set of objects and relationships. For each set of objects represented in this ontology, there is a description of its contents using regular expressions[3] and keywords. Therefore, an extraction ontology works as a sort of guide for the automatic creation of wrappers that perform extraction in a specific domain of interest. When applied to Web pages, it identifies objects and relationships, and associates them to elements of a conceptual model instance for that domain. Using this approach, the semantic meaning of each known term can be also recognized and extracted [17]. Martins et al. [30] use ontologies not only for the recognition and extraction, but also for the disambiguation of references to places.

Himmelstein [24] discusses the rapid growth of *local search*, a kind of geographically oriented search which focuses on the immediate geographic vicinity of the user, such as the city of neighborhood the user lives in, and explains why this subject is attractive to commercial and research sectors. Some commercial search tools have recently started offering geographic search capabilities. Such features allow users to locate places of interest near a given address and navigate the corresponding Web sites. Services such as Google Maps use Yellow Page business directories to retrieve information within a specific distance from a location. An online local search uses addresses for efficient proximity estimation.

An efficient geocoding strategy must exist to generate coordinates from textual addresses supplied by a Yellow Pages directory, or found within the text of a Web site. The quality of geocoding varies according to the strategy that is employed [49], and in turn the strategy depends on the quality and type of available addressing data. Such data may be found as numbering ranges over street centerline segments (also known as *street geocoding*, based on the TIGER file approach [47]), as numbers associated to land parcel centroids [35], or as numbers associated to individual buildings [11]. The last strategy has been developed as a response to addressing problems that are common in large cities of emerging countries, such as Brazil, due to the common occurrence of problems such as ambiguous street names and irregular numbering, which often rend commercial geocoding software useless. Furthermore, geocoding can use additional information to disambiguate and to approximate locations, such as indirect references to places (postal codes, building names, and telephone area codes).

Our approach differs from the ones mentioned previously, since it focuses on the local Web, i.e., pages concerning a given urban location. The approach is based on an ontology that has been designed to facilitate the process of recognizing, extracting, and geocoding partial or complete references to places embedded in text. It combines the extraction ontology approach described earlier with advances on urban gazetteers [42] created from data available on the Web and geocoding techniques [11]. This ontology, called OnLocus, is described in Section 4.

3 Geospatial evidence in web pages

Geographic meaning can occur anywhere on a page, making geographic context recognition a complex task. If we want to associate Web pages to places in a meaningful

[3] Regular expressions are constructs that specify a pattern used for matching character strings, usually employed in text processing [1, 19].

way, approximating what humans do and allowing for space-based navigation and selection of Web pages, we must be able to recognize such evidence and associate it with locations. This requires much effort towards understanding the semantic context of a page, and interpreting natural language expressions found in text.

References to places in Web pages can be direct or indirect. *Direct references* are usually mentions to place names, complete postal addresses, and sets of geographic coordinates. References to place names and addresses require additional data to be locatable, i.e., translated into a set of coordinates. *Indirect references* provide means to infer an approximate location from numeric or alphanumeric codes, such as postal codes and telephone area codes, or from expressions that indicate relationships to other places, which are directly referenced (for instance, "The hotel is two blocks from Times Square"). The association between indirect references and locations can change in time, as in the case of new postal codes. There can also be vague references, such as "South of France", for which no definite boundaries exist, and which vary according to individual perception [26].

Even though many works in the literature use place names as the main geospatial evidence within a page [25, 29, 40], alternative geospatial evidence in urban areas remains much less explored [4]. Indirect references, including postal addresses, phone numbers, postal codes, and landmarks, can be very helpful for both the disambiguation of references found in text, and to facilitate the determination of a geographic location corresponding to each reference. Furthermore, indications contained in indirect references can lead to useful approximate delineations, in an attempt to imitate human spatial reasoning [3].

We propose a three-phase process for recognizing geographic evidence in Web pages. The first one, *extraction*, is supported by extraction ontologies and aims at selecting relevant Web content. The second phase, *recognition*, corresponds to isolating references to places embedded in text and includes dealing with ambiguity. Gazetteers are used to support the recognition phase. Finally, the *location* phase obtains locations from the place descriptions previously recognized, using positioning data from gazetteers or from spatial databases. Each of these phases will be described in greater detail in the next subsections.

3.1 Extraction

Ontologies can provide semantic support for extracting and structuring semi-structured Web data [18]. As previously mentioned, extraction ontologies are characterized by their ability to recognize and classify value strings, especially in semistructured natural-language text. When applied to a document such as a Web page, an extraction ontology is able to identify objects and relationships that are associated to object sets and relationships sets within the conceptual model. As a result, the extraction ontology provides the necessary knowledge to wrap the page so that it becomes understandable within the conceptualization framework it encompasses.

Our interest lies in recognizing references to places in an urban context, as usually required by local searching, so we must create an *urban extraction ontology*. Such an ontology must describe rules for identifying elements within its domain that are present in Web pages, from which regular expressions and keywords indicating geographic context can be put together. Such rules must approximate the way people reason about the urban space, and the way they recognize references to places in text. In this work, we propose an urban place ontology as a resource to facilitate the extraction of geographic context from Web pages. Such an ontology allows associating a (possibly approximate) location for each geospatial evidence found, and permits recognizing and interpreting local place terms.

3.2 Recognition

The next phase involves the recognition of terms and expressions as place names. This can be achieved in a few different ways. Candidate terms and expressions can be compared to a gazetteer, a dictionary of place names. Alternatively, parsing and pattern matching techniques can be used to recognize structured references to places, such as postal addresses or telephone numbers.

Gazetteers such as Alexandria[4] and GeoNames,[5] online tools modeled after traditional dictionaries of place names, provide information elements to recognize references to places [21, 23]. Gazetteers manage place names in a global scale, including names of cities, rivers, mountains and other geographic features, associating each name to a geographic footprint. Recognizing place names with the aid of gazetteers is, however, limited by the quality and level of detail of gazetteer data, and by problems such as place name ambiguity.

According to Fu et al. [20], current gazetteers share limitations that keep them from being used more intensively in GIR. First, most spatial relationships are not coded in gazetteers. Second, generic relationships between object types are normally disconsidered, since they exceed what gazetteers are designed to achieve, but as a result the potential use of gazetteers with geographic ontologies is limited. Furthermore, geographic feature properties are defined using a simple representation or *footprint*, which usually lacks significant geometric details. Third, gazetteers associate names with such footprints, but lack support for fuzzy or semantically imprecise locations, such as "Southern California" [3, 26]. Considering the importance of recognizing place names in text, such limitations need to be addressed, and a new generation of gazetteers must be created [21], not simple organized listings of geographic names, but as tools to provide stronger support to GIR-related activities [50]. We add to those limitations the lack of intra-urban place names used by urban dwellers as points of reference for location or navigation, such as streets, neighborhoods, landmarks, references, monuments, and so on [42]. As a result, the current use of gazetteer data to recognize references to places must be accompanied by some sort of disambiguation technique.

Postal or street addresses can also be recognized, providing some of the most adequate geospatial evidence for local search applications. Addresses incorporate a time-proven system for locating a place based on a formal description. Addresses are practically universal in urban areas and embed the idea of narrowing down the search for a given place using a hierarchy of information items: country, state or province, city, neighborhood, street, building number [11]. Since postal addresses are usually presented in a standardized way, it is possible to recognize them in text from a previously determined template.

Although addresses are reasonably well-studied, especially for GIS purposes, the lack of a universal standard for addresses still complicates their recognition [31]. Address formats vary widely among countries, even though postal address recognition can be well established within each country [34]. Variations in Web page address elements, such as abbreviations, punctuation, and line breaks, complicate creating an address parser [39]. There may have some parts missing, such as the country [11, 31]. In such cases, indirect references serve as additional evidence to determine or to infer the missing pieces. For instance, if the address is incomplete but there is a postal code, it is possible to infer an association to a specific part of a country. Ground line phone numbers also carry location information implicitly, since numbering is usually organized according to geographic

[4] http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp
[5] http://www.geonames.org/

criteria. A complete telephone number identifies country code, area code, and city location. Since most traffic is local, however, phone numbers often omit country and area codes. Recognizing phone numbers in Web pages also requires precautions to avoid confusion with other numeric data, such as serial numbers. Since there exists a wide variation in number separators, such as dashes, parentheses, and blanks, a parser for phone numbers must be flexible enough to accommodate variations.

In the absence of addresses, sometimes references to locations can be indirectly recognized from natural language expressions that are semantically associated to place descriptions or wayfinding instructions [6, 13]. These expressions usually relate a relatively less-known place to another, more widely known one (a *landmark* of some kind), using one of many expressions that indicate spatial proximity, containment, or coincidence. We call such expressions *positioning expressions* [13], and will characterize them in greater detail in Section 4. Working with positioning expressions is a form of *qualitative spatial reasoning*, which implies the need for some tolerance as to imprecision and mismatches, since many times there is no way to associate concepts with precise locations or geographic objects. If that is the case, the recognition process must fall back to other evidence of spatial location contained in the page or use such additional information to infer a more general (therefore less precise) location from what is available.

3.3 Location

Once a reference to a place is recognized, the next step is to try to determine an actual location, which will then be associated to the source of the reference. In the case of direct references, we need to translate a place name or an address into a coordinate pair. This is achieved using footprint data from a gazetteer or performing a process known as *geocoding*, the determination of a location from a description, usually associated with the interpretation of addresses. In the case of indirect references, an estimation of the location can be made considering an interpretation of a positioning expression. Next subsections will present location techniques for each case.

3.3.1 Location of direct references

Once a place name has been recognized and disambiguated, a simple query to a gazetteer can return its footprint. Even though some places can be quite large spatially, usually a single coordinate pair will be provided (for instance, *New York City* is at 40° 42′ 51″N 74° 0′ 21″W).

In the case of postal addresses, some form of geocoding is required. Geocoding is the process that determines coordinates based on alphanumeric data, such as textual descriptions. The address is initially parsed into its fundamental components, such as street name and number, neighborhood name, city/state name, and possibly a postal code. These elements are then used in two stages: *matching* and *locating*. In the matching stage, a correspondence is established between the identified address and a geographic entity from the gazetteer (such as a street, neighborhood, or city). In the locating stage, geographic coordinates are associated with the address. The geocoding process requires the existence of georeferenced addressing data, possibly down to point-georeferenced individual addresses or street segments associated to numbering ranges. Results can be *exact*, when the extracted address corresponds exactly to an address available in the gazetteer, or *approximate*, when the location is estimated from nearby elements, such as the closest building number on the street or the street segment whose numbering range includes the provided building number.

If neither an exact nor an approximate location can be determined, a *general location* can often be established using information such as the neighborhood name, postal code, or city limits [11].

3.3.2 Location of indirect references

Indirect references comprise indications that allow people to approximate locations, as in the case of postal codes or telephone area codes, and expressions that indicate a location in relation to other places. In the first case, it is necessary to establish a correspondence between a code and the area it serves, an approach that can be directly supported by spatial databases. In the second case, some natural language interpretation is required.

Natural language offers many different ways to indicate the location of something, for instance, the positioning expression that indicates that "a shop" is located "at the third floor of" "Big Mall". Therefore, if we can determine the location of Big Mall (a direct reference to a place), we are sure to get a good approximation of the shop's location as well, and the mall's address is more widely known. Since indirect references embed a direct reference to a place, the techniques presented in the previous section apply. We will cover here techniques that allow the estimation or approximation of a location based on expressions that indicate spatial location, i.e. *positioning expressions*.

A positioning expression is a natural language expression formed by the connection of two components—a subject and a landmark—using a third component—an expression that denotes a spatial relation. Therefore, a positioning expression is a triple <S, SR, L>, where S is the subject, SR is the spatial relation, and L is the landmark. The *subject* is an entity, usually a place ('someone's house', 'a bus stop', 'a tourist attraction'), event ('a burglary', 'a public gathering'), or product ('pizza', 'tyres')—whose approximate location of occurrence is textually described by a positioning expression. The spatial *relation* is a natural language expression that indicates the spatial connection between the subject and the landmark.[6] The *landmark* is represented in text by a place name, thereby corresponding to a direct reference to a place.

The interpretation of positioning expressions is based on the cognitive meaning of landmarks in human spatial orientation. In a previous study [13], we distinguished between two main groups of expressions that indicate spatial location: expressions that denote spatial containment or coincidence (for instance, *in front of*, *beside*, *next to*, *inside*), and expressions that indicate proximity (for instance, *close to*, *a few minutes from*, *within walking distance*). In the case of containment and coincidence, a location can be determined from the landmark's position. In the case of proximity, determining an actual location is more difficult, but the immediate vicinity of the landmark would serve as a rough first approximation. From experiments performed on four million pages from the Brazilian Web, we have been able to establish an approximate distance that corresponds to each expression that indicates proximity [13]. We realized that some expressions convey spatial proximity more emphatically than others. However, regardless of the exact distance, there is a semantic equivalence between such expressions, around the concept of proximity.

[6] Previous works [15] have precisely characterized spatial relations using point-set and other mathematical concepts, and named each resulting relation. While these names are traditional in the GIS community, not everybody uses them in natural language, and their interpretation remains ambiguous for our purposes.

4 OnLocus: an extraction ontology of urban places

Based on the needs for the extraction, recognition, and location of geographic evidence from text, we propose OnLocus, an ontology of urban places. OnLocus provides a hierarchical and semantic location structure on urban geographic spaces and their associations, and therefore can be classified as both a geographic and an extraction ontology [17] and is domain dependent [8]. It is a geographic ontology because it semantically describes concepts related to urban location; it is domain dependent because it defines a group of concepts that are commonly found in most urban communities. Furthermore, according to the classification proposed by Spaccapietra et al. [43], OnLocus includes two ontology types: taxonomic and descriptive. The taxonomic ontology is called "space taxonomic ontology", because it explores the hierarchical structure of urban space. The descriptive ontology considers reference points and various forms of place descriptions meaningful to the local community. Figure 1 presents a schematic and simplified version of OnLocus, included here to provide an overview of the ontology. The notation used in the presentation of OnLocus is described in Fig. 2. Concepts are represented by rectangles and relationships by lines. To improve semantic representation, some graphic primitives were introduced to establish a distinction among different types of relationship. A more formal and thorough encoding of the ontology has been developed using Protégé,[7] but we will refrain from showing it directly due to space limitations.

Place is the main concept of OnLocus. Every instance of *Place* can be referenced using a description, which OnLocus defines as a *place descriptor*. Place descriptors include addresses, place names, and positioning expressions. Places are specialized into *Territorial Division* and *Landmark*. Each of these concepts is further specialized, as described next.

4.1 Place

The *Place* concept represents space descriptions related to locations, which can be identified, or referred to, in various ways. A place is usually represented in a Geographic Information System (GIS) by its geometry and topology, but for OnLocus it is a concept that includes a cognitive aspect, reflecting how people think geographically and use geographic information in a daily basis. This decision reflects our observation that people often refer to places in regular speech using approximate references and connections to more widely-know places, as in the expressions "near the airport", "down the road from the conventions center", or "in the region of the monument". Such place descriptors are not recognizable by regular GIS functions or queries, but are very important in the context of geographic information retrieval.

Each instance of *Place* is associated to its spatial representation. This representation uses geometric concepts usually associated to GIS, such as points, lines, and polygons. Line and polygon geometry can be approximated through the use of minimum bounding rectangles (MBR). Places are generalizations of descriptions from two other concepts: territorial divisions and landmarks (Fig. 1).

4.2 Territorial division

OnLocus defines a *Territorial Division* when an area is recursively subdivided into a set of smaller ones. The territorial division hierarchy is reflected in the mereologic relationships

[7] http://protege.stanford.edu

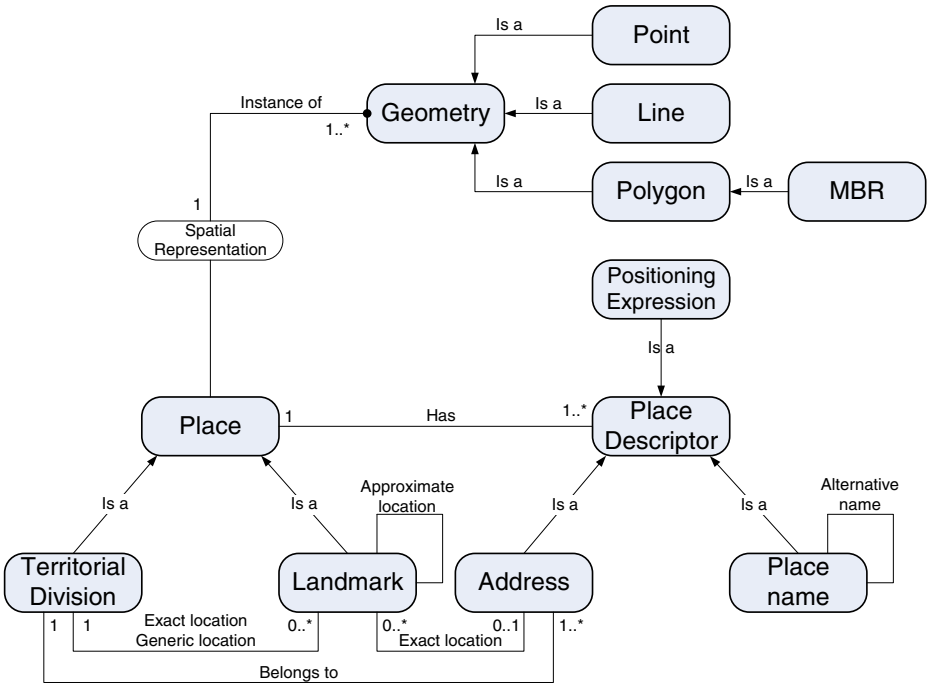


Fig. 1 OnLocus simplified schema

among concepts it embodies, such as city, state, area, and country. The concept of territorial division is related to all kinds of subdivision of space into regions. Such divisions are commonly used in political and administrative divisions, such as countries, states, and municipalities, and therefore are usually hierarchical. Figure 3 presents an instance of this concept for Brazil, showing subdivisions of a country's territory arranged as a hierarchy. The finest granularity included in OnLocus corresponds to subdivisions of the urban space. Notice that, along with divisions that are based on political boundaries, others are defined by more practical and operational concerns, such as postal code areas and telephone code areas. In fact, many other kinds of divisions can exist, based on physical characteristics (for instance, watersheds or terrain slope ranges) or on other kinds of classification (such as vegetation or soil type). For the sake of clarity, and for keeping up with the objectives of this article, we will restrict ourselves to divisions that can be relevant for urban applications.

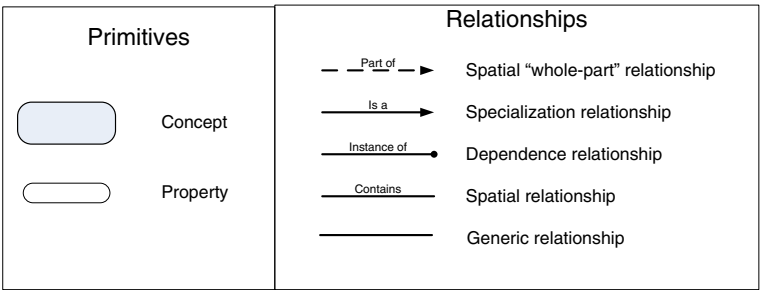


Fig. 2 OnLocus schema primitives

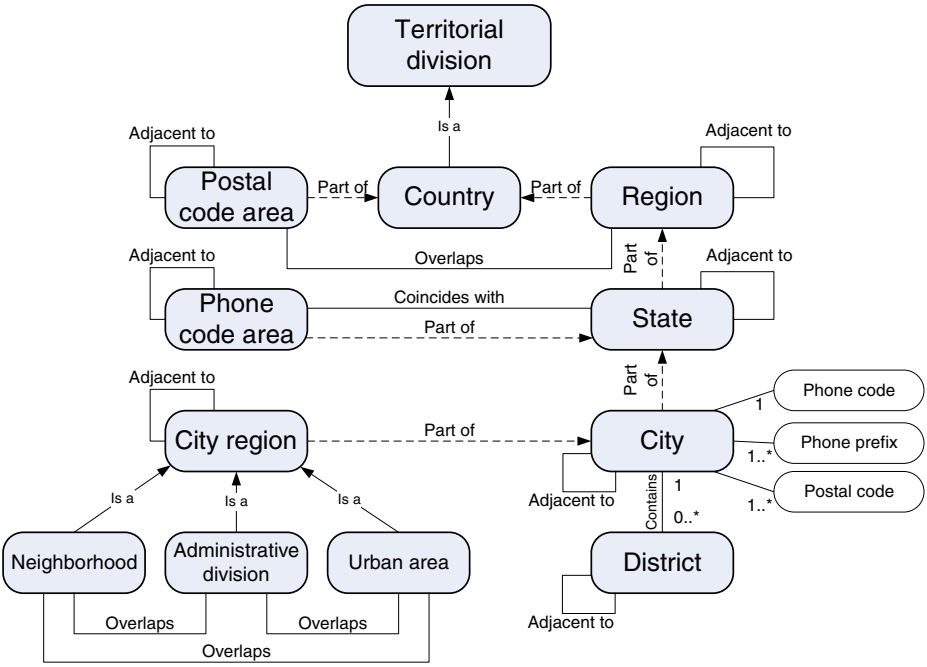


Fig. 3 Example of the Brazilian Territorial division in OnLocus

4.3 Landmark

A *landmark* represents a specific place, such as an urban or environmental point of reference, which is known by enough people, so that it can be used for spatial orientation. In OnLocus, this concept comprehends several other types of places that are used as a reference for spatial navigation or orientation, such as tourist attractions, parks, or museums. Figure 4 presents several concepts associated to urban places. Landmarks are specialized beyond the concepts presented in the figure. For instance, a "reference building" can be a school, hospital, or temple, while a "culture and leisure" landmark can be a museum, theater or park. This kind of definition is extensively detailed in other ontologies, such as OpenCyc[8] and SUMO,[9] as well as in gazetteers such as TGN,[10] and can be reused by OnLocus. However, there are cases in which concept reuse is not entirely possible, due to differences in hierarchical position and to the use of concepts in the extraction ontology that characterizes OnLocus and differentiates it from the mentioned ontologies.

4.4 Place descriptor

The concept of *place descriptor* defines the various ways people use to refer to a place. Places are most often recognized by their descriptors, which correspond to an *address*, a *toponym* or *place name*, or a *positioning expression*. A place can be referred to by more

[8] http://www.opencyc.com
[9] http://www.ontologyportal.org
[10] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

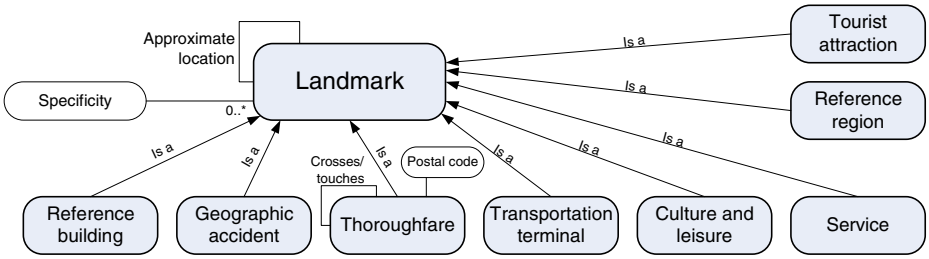


Fig. 4 Landmarks in OnLocus

than one descriptor; for instance, a tourist attraction can be recognized by name (The White House), but it can also have a corresponding address (1600 Pennsylvania Ave., Washington DC), and can be located using a positioning expression (close to The Mall). Figure 1 schematically presents place descriptors and their specializations. OnLocus associates each of the place descriptor types to specifications, such as patterns and regular expressions, that will be used to recognize references to places in Web pages. The next subsections will present each specialization of place descriptors in more detail.

4.4.1 Address

In OnLocus, the address concept includes several elements, which vary according to the purpose for which the address is used and to the place to which the address is associated. OnLocus divides addresses into three parts: *Basic Address*, *Complement*, and *Location Identifiers* (Fig. 5). The complement is the set of additional addressing information beyond the street name and number, including elements such as apartment number, neighborhood name, suite number and others. OnLocus considers phone numbers as part of the addressing components. Phone numbers, especially area codes, are instrumental in determining an approximate location and in resolving ambiguities.

Addresses can be *complete*, *incomplete*, or *partial*. A complete address includes all components generally associated to postal delivery, such as thoroughfare identification, building number, complement, city, state, and postal code. Depending on the elements that are missing, incomplete addresses can be ambiguous, since the same street name can occur in different cities, or imprecise, since existing elements can be insufficient to allow for an exact placement. Resolving such ambiguity requires some indication of context, such as a phone number with area code, leading to a unique place [11, 12]. A partial address includes only location identifiers, with which only approximate locations can be determined.

Language and local culture must be observed in address recognition. Even though addresses are used worldwide and are formed of essentially the same components, the sequence in which these components appear varies among countries (Fig. 5). The parser must correctly identify the order of address components for extraction [11]. OnLocus may contain versions adapted to any possible address variation.

```
Address( BasicAddress(StreetType, StreetName, BuildingNumber), Complement,
LocationIdentifiers(PhoneNumber(AreaCode,Number),PostalCode,
CityState(City, State)) )
```

Fig. 5 OnLocus address structure (Brazilian style)

Figure 6 shows an example of the expressions that are included in OnLocus as a resource for recognizing addresses. The expressions are shown in EBNF (Extended Backus-Naur Form) [38], and present a translated and adapted version of a Brazilian-type basic address, following the structure presented in Fig. 5. In the first definition, *ADDRESS*, "BR" refers to the prefix assigned to all federal highways in the country. In *STREET_TYPE*, the idea is to record all possible alternatives of thoroughfare types and their abbreviations, including capitalization variations. *PREP* and *S1* act as separator stopwords, the first one corresponding to common prepositions that are included in street names that are composed of more than a single word (for instance, "Avenue of the Americas"). *IDENT* and *N* are respectively address and number prefixes.

4.4.2 Place name (toponym)

The *toponym* concept represents the names people give to places. A place can be known by several different names and is, therefore, subject to ambiguity. Likewise, the same name can be associated to many different places. In Fig. 1 the association of multiple names to a place, and of multiple places to a single name, are indicated by recursive relationships.

The extraction of place names is not usually performed using simple regular expressions, because of the need to disambiguate and to distinguish between place names and other words. This is done with the help of gazetteers, although most gazetteers do not include disambiguation tools or information resources, as discussed in Section 2. For that purpose, we used information from Locus [42], a non-conventional gazetteer developed by our group that includes intra-urban names and indirect references, to recognize references to places in text. We are currently enhancing Locus in order to achieve better results in place name disambiguation for GIR. For the experiments presented here, most place names were

Basic Address pattern

```
<ADDRESS>::= [<IDENT>] (<STREET_TYPE> <STREET_NAME> | ("BR" | "BR.") {0-9}+)
             [<S1>] <NUM>

<IDENT>::= ("Address:" | "ADDRESS:"| "ADDR:"| "Addr:"| "ADDR.:"| "Addr.:"|
           "Location:"| "Place:");

<STREET_TYPE>::= ("Street" | "STREET" | "St." | "St" |
                  "Avenue" | "AVENUE"| "Av" |"AV" | "Av."|"AV." |"Ave"|"AVE" |
                  "Highway" | "HIGHWAY" | "Hwy" | "HWY" | "Hwy." |"HWY." |
                  "Square" | "Sq"| "SQ"| "Sq."| "SQ."|
                  "Parkway" | "PARKWAY"| "Pkwy."| "PKWY."| "Pkwy"| "PKWY"|
                  "Road" | "ROAD" | "Rd"| "RD"| "Rd."| "RD."|
                  "Drive" | "DRIVE" | "Dr" | "DR" |"Dr." | "DR.");

<STREET_NAME>::= ({1-9} <PREP>  {(A| B |...|Z) {a-z}}+ |
                  {{(A| B |...|Z) {a-z}}+  <PREP>  {(A| B |...|Z) {a-z}}+}+ |
                  {(A| B |...|Z) {a-z}}+ {1-9}|
                  {{(A| B |...|Z)}+ {a-z}}+|
                  {1-9}+ ) ;

<PREP>::= ("of" | "on" |  "at"|  "in" | "of the" | "   ")  ;

<S1>::= (" , " |  " - " | "    ");

<NUM>::= ([<N>] {(0|1|...|9)}+ " . " (0|1|..|9) (0|1|..|9) (0|1|..|9) {a-z} |
         {A-Z})  |
         [<N>] {(0|1|...|9)}+ ({a-z} | {A-Z}) ) ;

<N>::= ("n." | "N." |"n" | "N" | "#" | "Mile Marker"| "Number" | "number");
```

Fig. 6 Basic address pattern in EBNF notation (Brazilian style)

derived from either positioning expressions (as described in the next section) or from postal addresses. As we will show in the next sections, it is easier and more efficient to use indirect or structured references such as addresses as unambiguous references to places, but recognizing place names as such within natural language text remains an important task.

4.4.3 Positioning expression

A *positioning expression* is a semantic construction in natural language that is frequently used by people to indicate or to approximate the location of something in relation to some other place whose location is more widely known. A positioning expression is defined as a pair <*spatial relationship, place name*> [13] e.g., <close to, Big Mall>.

Some natural language expressions that denote spatial relationships are listed in Section 3.3.2. Table 1 presents some of the most popular positioning expressions in Portuguese, as found in the Brazilian Web [13], along with their meaning in English. Figure 7 presents an example of a pattern for a positioning expression, once again translated and adapted from Portuguese. The example has been formulated from a regular expression, which is included in OnLocus.

4.5 Relationships

OnLocus considers predefined binary relationships, both conventional and spatial. Conventional relationships occur between concepts and between properties. **Specialization** is a conventional relationship used to create a hierarchy in which generic concepts are specialized into more specific concepts [41, 46]. **Aggregation** is a relationship used to indicate the composition of a concept from a set of properties. It is usually denoted as a "part-of" relationship, meaning that the more general concept includes all properties of the more specific concepts. **Precedence** is a structural constraint between properties in an aggregation. It defines a standard sequence for the properties, in order to group them according to an expected set of rules. The resulting sequence causes a concept to be recognized by its properties. For instance, in Brazilian addresses it is expected that thoroughfare types come first, then a thoroughfare name, and then a building number. For U.S. addresses, the correct order would be the building number, then the thoroughfare name, then the thoroughfare type. **Dependency** is a relationship that is characterized by an "instance-of" association, which happens when the value of a property is an instance of another property. For instance, in Fig. 1 the *spatial representation* property of the *place*

Table 1 Natural language expressions denoting containment or coincidence

In Portuguese	In English (semantic meaning)
Dentro de	Inside, in, into, within
No coração de	In the heart of (in the middle of, downtown)
No n-*ésimo piso de*	In the *n*-th floor of
No n-*ésimo nível de*	In the *n*-th level of
Na praça de alimentação	At the food court
Em cima de	Above (upstairs, uphill, up the street, further along the street)
Embaixo de	Below (downstairs, downhill, down the street, further along the street)
No / na / em	In, at

Example of positioning expression pattern

```
<POS_EXPRESS1>::= <SPATIAL_REL1> [<NUM>] [<PREP1>] <PLACE_NAME>

<SPATIAL_REL1>::= ("ONLY A FEW" | "only a few"| "a little more than"| "A LITTLE MORE
            THAN"| "LESS THAN"| "less than"| "near"| "NEAR"|  "almost"| "ALMOST"|
            "EXACTLY"| "exactly" |"AROUND" | "around");

<NUM>::= ( {(0|1|...|9)}+ " . " (0|1|..|9) {(0|1|..|9)} <N> |
         {(0|1|...|9)}+ <N> |
         {(A| B |...|Z) {a-z}}+ <N>|
         {a-z}+ <N>) ;

<N>::= ("mile"| "MILE" |"Mile" |"Miles" | "MILES" | "miles" | "kilometers" |
        "KILOMETERS" | "Kilometers" | "KILOMETER" |"Kilometer" |"kilometer"| "Km" |
        "km" | "Mi" | "mi" );

<PLACE_NAME>::= ({{(A| B |...|Z) {a-z}}+  <PREP2>  {(A| B |...|Z) {a-z}}+}+ |
                 {(A| B |...|Z) {a-z}}+ |
                 {{(A| B |...|Z)}+ {a-z}}+ );

<PREP1>::= ("for" | "from" | "   ")

<PREP2>::= ("of" | "on" |  "at"|  "in" | "of the" | "   ")  ;
```

Fig. 7 Example of positioning expression pattern in EBNF notation

concept can include instances of the concept *geometry*. **Generic** relationships represent associations between instances of different concepts. A generic relationship is usually expressed by a verb, chosen according to the semantics of the relationship. For instance, a *place* "has" a *place descriptor*.

Spatial relationships occur only between concepts and belong to three basic types: topological, "whole-part", and location. Clementini et al. [10] present five distinct **topological relationships**, which occur between point, line, and polygon geographic objects. For the purposes of this work, the formal definition of topological relationships is not as important as the meaning people assign to expressions that denote such relationships in natural language. The five basic topologic relationship types are *touch*, *in*, *overlap*, *cross*, and *disjoint*. We also define *adjacent to* and *coincide with*. These relationships are basically equivalent to the ones defined by Egenhofer and Franzosa [15] using the 4- and 9-intersection matrixes. Observe that, even though the relationships receive names that correspond approximately to their meaning in natural language, people use many other terms that refer to the same relationships. For that reason, OnLocus does not use the *disjoint* relationship. Furthermore, the *coincide with* relationship is used whenever the same spatial subdivision is used as a reference for two or more concepts and the *adjacent to* relationship is used with planar subdivisions [5], in order to characterize a variation of the *touch* relationship. Directional relationships (i.e., to the North of, West of) [22] also fit this category, with a similar treatment. We will not expand on directional relationships here due to space limitations.

"Whole-part" relationships are mereologic relationships between concepts. A concept denoted as C_j is *part-of* a concept C_i, if C_i has C_j as one of its parts or if C_j is a *part-of* C_i such that $C_j \subset C_i$. Our approach to **location relationships** is inspired by the work of [9], who propose the classification of terms into *spatially unstructured* (e.g. containing indirect references to places) and *spatially structured* (e.g., which use unmistakable references to positions, such as coordinates). In our approach, a location relationship can represent three types of location: *exact*, *approximate*, and *generic*. Such relationships imply that the position of an object can be determined relatively to the position of another one, whose location is previously known. *Exact location* is a binary relationship usually defined

between geographic objects and known spatial regions (usually polygon objects). *Approximate location* is a binary relationship between geographic objects, in which the position of an object is approximately determined from the position of another. *Generic location* is a binary relationship between spatial objects and space subdivision hierarchies. When obtaining the exact position is not possible, a generic location is attempted, in order to determine that a spatial object is located "somewhere" inside a region. In generic locations there is no assurance of precise location, just the establishment of a relationship between an object and the region.

4.6 Using OnLocus for the extraction of geographic evidence

Figure 8 presents the main steps of a process for recognition, extraction and geocoding of geographic evidence from Web pages based on OnLocus. Initially, Web pages are collected and pre-processed, applying the usual cleanup procedures used in information retrieval, such as normalization of the set of HTML delimiters, removal of consecutive spaces and removal of accent marks. After collection and pre-processing, the pages move on to the recognition and extraction of potential geospatial references, such as addresses, postal codes, and phone numbers. From the definitions and structure of OnLocus, as presented in the previous subsections, we specified a number of patterns for the recognition of geographic evidence from text. Patterns for the recognition of basic addresses, postal codes,

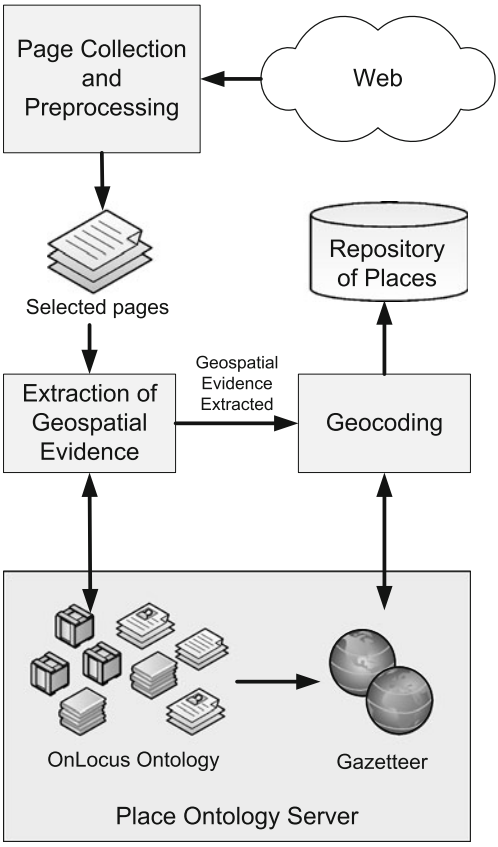


Fig. 8 Ontology-driven extraction

city names, states, and telephone numbers were created using EBNF, as shown in Figs. 6 and 7. Each of these patterns was transformed into a regular expression, which were then implemented as a Perl script.

As a result of the process, a *repository of places* is formed. This repository contains, for each extracted and validated evidence, the URL in which it was found, the pattern used in the extraction, the extracted terms, the initial and final position of the terms on the page, the city and state names, a set of geographic coordinates corresponding to the minimum bounding rectangle (MBR) of the city's limits, and, if possible, the geographic coordinates associated to the address.

5 Experimental evaluation

The OnLocus ontology has been employed in several experiments of data extraction from Web pages, using the WBR05 collection, comprising over 4 million pages from the Brazilian Web, crawled in March 2005 [33]. This collection is representative of the Brazilian Web.

Our experiments were designed to determine the efficiency of the retrieval of positions from addresses embedded in Web pages. Initially, 17 different address patterns were used, in order to verify which ones would be more useful for retrieval. Each pattern corresponds to a possible combination of address components in which addresses are usually found in text. Six of the patterns stood out in the results from this preliminary evaluation [6], based on their extraction capabilities: *PhoneNumber*, *BasicAddress* + *CityState* + *PostalCode*, *BasicAddress* + *PhoneNumber*, *BasicAddress* + *CityState*, *BasicAddress* + *PostalCode*, and *PostalCode*. The good performance of these patterns can be explained by the fact that they correspond to the more usual and conventional ways in which addresses appear in text, a kind of knowledge that has been included in OnLocus. We also used information from the Locus non-conventional gazetteer [42] to recognize references to places in text.

These patterns were used over the WBR05 collection. Matches were recognized in 603,798 pages, or 14.8% of the number of pages in the collection, thus confirming previous results [24, 31]. We recognized 2,137,601 occurrences of the patterns, over 3.5 occurrences per page. Table 2 shows the distribution of matches among the patterns. Notice that *PhoneNumber* accounted for more than half of the matches, and was found in most of the pages. Most addresses include either the city and state, or the postal code, but only a few include both. *PostalCode* was also a frequently found pattern, accounting for almost a quarter of the occurrences.

We geocoded the patterns we were able to extract, using processes and algorithms described in more detail by Davis Jr. and Fonseca [11], in order to establish how good a location we can expect to obtain.

In order to verify the automatic results shown in Table 2, we randomly selected 385 extractions for each pattern for manual inspection. The inspection of the geocoding results showed that, for the patterns including a phone number, failures resulted from outdated numbers, with a wrong area code or presented in an unusual format. In patterns including city/state most problems were in the recognition of the city name. The main issues include (1) lack of a separator between a neighborhood name and the city name; (2) cities that have the same name as the state they are in; (3) landmark names used where a city name was expected; (4) abbreviations in the city name. Overall, from the 2,310 extractions (6 patterns times 385 pages per pattern) manually inspected, 80% were found to be correct, and were correctly geocoded. Table 3 details the results for each pattern. The worst performance

Table 2 Summary of extraction from WBR05

Pattern	Pages including the pattern		Occurrences of the pattern	
	Number of pages	%	Number of occurrences	%
PhoneNumber	505,189	83.7%	1,083,913	50.7%
BasicAddress + *CityState* + *PostalCode*	24,475	4.1%	34,832	1.6%
BasicAddress + *PhoneNumber*	55,244	9.1%	99,297	4.6%
BasicAddress + *CityState*	155,063	25.7%	217,274	10.2%
BasicAddress + *PostalCode*	154,761	25.6%	231,406	10.8%
PostalCode	285,999	47.4%	470,879	22.0%
Total	603,798 (*)	(*)	2,137,601	100.0%

(*)—The same page can contain more than one occurrence of a pattern or patterns

occurred in the *BasicAddress* + *CityState* pattern, in which the extraction pattern captured many other groups of words with the same layout, but different meaning. Many mistakes occurred in the geocoding, although the extraction was correct, showing that the geocoder's performance is dependent on the quality of the contents of reference tables for streets and addresses [11].

Table 4 shows the results of city-level geocoding based on three components of the recognition patterns: *PhoneNumber*, *PostalCode*, and *CityState*. Results show that the *PostalCode* provided the most effective way to obtain approximate locations, with over 97% of success. Phone numbers are also a good indication, with a geocoding success rate of about 80% (the remaining 20% lack the area code, so they are potentially ambiguous with phone numbers from elsewhere). Success in recognizing *CityState* was the lowest. This can be explained by the fact that this component offers more possibilities for format variation and misspellings, therefore leading to pattern matching errors. In the pattern in which both *PostalCode* and *CityState* are available, using the postal code also led to better results.

Notice also that geocoding was more successful in the cases in which basic address data are available. This can be explained by the fact that addresses provide additional disambiguation information. Of course, obtaining a more precise location requires interpreting the elements that form the *BasicAddress* component. Since the city is

Table 3 Geocoding results per pattern

Pattern	Extractions	Geocoded extractions	Not Geocoded	
			Incorrect Extractions	Incorrect Geocoding
PhoneNumber	385	262	3	120
BasicAddress + *CityState* + *PostalCode*	385	385	0	0
BasicAddress + *PhoneNumber*	385	269	1	115
BasicAddress + *CityState*	385	193	189	3
BasicAddress + *PostalCode*	385	380	0	5
PostalCode	385	359	22	4
Total	2,310	1,848 (80%)	215 (9%)	247 (11%)

Table 4 Number of addresses extracted and geocoded

Pattern	Extracted	PhoneNumber Geocoded (%)	PostalCode Geocoded (%)	CityState Geocoded (%)
PhoneNumber	1,083,913	865,966 (79.89%)		
BasicAddress + *CityState* + *PostalCode*	34,832		34,600 (99.33%)	23,757 (68.20%)
BasicAddress + *PhoneNumber*	99,297	80,884 (81.46%)		
BasicAddress + *CityState*	217,274			167,488 (77.07%)
BasicAddress + *PostalCode*	231,406		229,851 (99.33%)	
PostalCode	470,879		453,380 (96.28%)	
Total	2,137,601			

successfully determined in most of the cases, the possibility of ambiguity in geocoding gets much lower.

The presence of a postal code suggests the existence of a complete postal address somewhere else in the page. However, the recognition of postal codes was more successful than the recognition of the pattern *BasicAddress* + *PostalCode* (22% vs. 11%, see Table 2). This difference shows how hard it is to correctly recognize postal addresses, and indicates that exploring indirect references such as the postal code or the phone number can lead to more successful results. However, finding a postal code after a basic address allows us to avoid confusion with other numerical data presented in a similar format.

6 Conclusions and future work

This paper focused on the local Web and presented an approach based on an ontology of urban places that allows recognition, extraction, and geocoding of geospatial evidence with local characteristics. It presented a method to identify the most common patterns for address extraction and a minimal set of patterns for the extraction of Brazilian addresses was obtained and validated experimentally using a collection of over 4 million Web pages. We described experiments evaluating the presence and incidence of urban addresses and positioning expressions in Web pages. Addresses and positioning expressions provided satisfactory support for local search applications, since they represented the physical location of services and activities found in Web pages, and because of the high success rate achieved in geocoding.

One of the main goals of geospatial evidence recognition is to allow the creation of mechanisms which enable search engines to perform local and proximity searches, without resorting to yellow page directories. The experiments presented here showed the feasibility of performing automated address extraction and geocoding to identify locations associated to Web pages. The combination of location identifiers with basic addresses improved the precision of extractions and reduced the number of false positive results.

Results indicate that, in many cases, we were able to obtain geographic references from Web pages that refer to intra-urban locations. Finding a relationship between a page and a city, or cities, is currently more efficient, which indicates that substantial improvements can derive from better and more generic ways to recognize and geocode addresses presented in irregular formats, as well as from better disambiguation techniques. We were able to benefit from the non-conventional features of Locus [42], a gazetteer that includes intra-urban

names and indirect references. Locus is currently evolving, in order to include additional disambiguation information and to expand inference possibilities based on spatial relationships.

The results presented in this article open perspectives for new types of useful applications which can simplify, improve, and enhance local Web searches. Future work includes expanding search engine queries using OnLocus to determine semantically-equivalent positioning expressions or related places, and developing new ways to navigate between pages considering their geographical relationships to the same or similar places. We are also considering the use of semantic annotations to indicate relationships between pages and places.

References

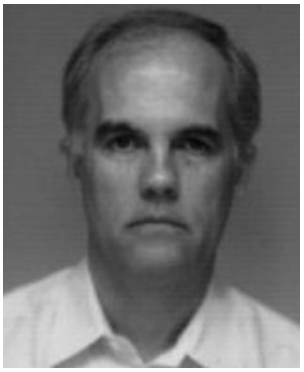
1. Aho AV (1990) Algorithms for finding patterns in strings. Handbook of theoretical computer science. In: van Leeuwen J (ed) Volume A: Algorithms and complexity. The MIT Press, pp 255–300
2. Amitay E, Har'El N, Sivan R, Soffer A (2004) Web-a-Where: Geotagging Web Content. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, pp 273–280
3. Arampatzis A, van Kreveld M, Reinbacher I, Jones CB, Vaid S, Clough P, Joho H, Sanderson M (2006) Web-based delineation of imprecise regions. Comput Environ Urban Syst 30:436–459
4. Borges KAV (2006) Use of an ontology of urban places for recognition and extraction of geospatial evidences on the web (in Portuguese). Belo Horizonte (MG), Brazil, Federal University of Minas Gerais
5. Borges KAV, Davis CA Jr, Laender AHF (2001) OMT-G: an object-oriented data model for geographic applications. GeoInformatica 5(3):221–260
6. Borges KAV, Laender AHF, Medeiros CB, Davis CA Jr (2007) Discovering geographic locations in web pages using urban addresses. Proceedings of the 4th ACM Workshop on Geographic Information Retrieval, Lisbon, Portugal, pp 31–36
7. Borges KAV, Laender AHF, Medeiros CB, Silva AS, Davis CA Jr (2003) The web as a data source for spatial databases. Proc. of the V Brazilian Symp. on GeoInformatics, Campos do Jordão (SP), Brazil: CD-ROM
8. Buneman P, Khanna S, Tan W-C (2000) Data provenance: some basic issues. FST TCS 2000: Foundations of software technology and theoretical computer science: 20th conference. New Delhi, India: p87
9. Casati R, Varzi AC (1996) The structure of spatial localization. Philos Stud 82:205–239
10. Clementini E, DiFelice P, van Oosterom P (1993) A small set of formal topological relationships suitable for end-user interaction. 3rd Symposium on Spatial Database Systems: 277–295
11. Davis CA Jr, Fonseca FT (2007) Assessing the certainty of locations produced by an address geocoding system. Geoinformatica 11(1):103–129
12. Davis CA Jr, Fonseca FT, Borges KAV (2003) A flexible addressing system for approximate urban geocoding. V Brazilian Symposium on GeoInformatics (GeoInfo 2003), Campos do Jordão (SP):CD-ROM
13. Delboni TM, Borges KAV, Laender AHF, Davis CA Jr (2007) Semantic expansion of geographic web queries based on natural language positioning expressions. Trans GIS 11(3):377–397
14. Ding J, Gravano L, Shivakumar N (2000) Computing geographical scopes of web resources. Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt: 545–556
15. Egenhofer M, Franzosa R (1991) Point-set topological spatial relations. Int J Geogr Inf Syst 5(2):161–174
16. Egenhofer MJ (2002) Toward the semantic geospatial web. Geographic Information Science 2002. McLean, Virginia, pp 1–4
17. Embley DW (2004) Toward semantic understanding—an approach based on information extraction ontologies. Proceedings of the 15th Australasian Database Conference, Dunedin, New Zealand, pp 18–22
18. Embley DW, Campbell DM, Jiang YS, Liddle SW, Lonsdale DW, Ng Y-K, Quass D, Smith RD (1999) Conceptual-model-based data extraction from multiple-record web pages. Data Knowl Eng 31(3):227–251
19. Friedl J (2002) Mastering regular expressions. O'Reilly

20. Fu G, Jones CB, Abdelmoty A (2005) Building a geographical ontology for intelligent spatial search on the web. Proc. of the IASTED Int'l Conf. on Databases and Applications, Innsbruck, Austria, pp 167–172
21. Goodchild MF, Hill LL (2008) Introduction to digital gazetteer research. Int J Geogr Inf Sci 22 (10):1039–1044
22. Goyal RK (2000) Similarity assessment for caardinal directions between extended spatial objects. Orono, Maine, University of Maine, p189
23. Hill LL (2000) Core elements of digital gazetteers: placenames, categories, and footprints. 4th European Conference on Research and Advanced Technology for Digital Libraries, pp 280–290
24. Himmelstein H (2005) Local search: the internet is the yellow pages. IEEE Comput 38(2):26–35
25. Jones CB, Purves R, Ruas A, Sanderson M, Sester M, van Kreveld M, Weibel R (2002) Spatial information retrieval and geographic ontologies: an overview of the SPIRIT project. ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, pp 387–388
26. Jones CB, Purves RS, Clough PD, Joho H (2008) Modelling vague places with knowledge from the web. Int J Geogr Inf Sci 22(10):1045–1065
27. Laender AHF, Borges KAV, Carvalho JCP, Medeiros CB, Silva AS, Davis CA Jr (2005) Integrating web data and geographic knowledge into spatial databases. Spatial databases: techniques, technologies and trends. In: Manolopoulos Y, Papadopoulos A, Vassilakopoulos M. Hershey Pennsylvania, USA, Idea Group Publishing, pp 23–48.
28. Larson RR (1996) Geographic information retrieval and spatial browsing. Geographic information systems and libraries: patrons, maps, and spatial information. In: Smith LC, Gluck M (eds). Urbana, IL, Un. of Illinois, pp 81–123
29. Manov D, Kiryakov A, Popov B, Bontcheva K, Maynard D, Cunningham H (2003) Experiments with knowledge for extraction. Proceedings of the Human Language Technology Conference Workshop on Analysis of Geographic, Edmonton, Canada, pp 1–9
30. Martins B, Silva MJ, Freitas S, Afonso AP (2006) Handling locations in search engine queries. Proceedings of the 3rd ACM Workshop on Geographical Information Retrieval (GIR 2006), Seattle, Washington, USA
31. McCurley KS (2001) Geospatial mapping and navigation on the web. Tenth International World Wide Web Conference (WWW10), Hong Kong, ACM, pp 221–229
32. Miller C (2006) A beast in the field: the google maps mashup as GIS/2. Cartographica Int J Geogr Inf Vis 41(3):187–199
33. Modesto M, Pereira Á Jr, Ziviani N, Castillo C, Baeza-Yates R (2005) A new portrait of the Brazilian Web (in Portuguese). Proceedings of the XXXII Seminar on Integrated Software and Hardware (SEMISH 2005), São Leopoldo (RS), Brazil, pp 2005–2016
34. Rhind G (1999) Global sourcebook of address data management: a guide to address formats and data in 194 countries gower
35. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL (2006) Geocoding in cancer research: a review. Am J Preventative Med 30(2S):S16–S24
36. Sanderson M, Kohler J (2004) Analyzing geographic queries. Proc. of the ACM SIGIR Workshop on Geographic Information Retrieval, Sheffield, UK, pp 1–2
37. Schockaert S, De Cock M, Kerre EE (2008) Location approximation for local search services using natural language hints. Int J Geogr Inf Sci 22(3):315–336
38. Scowen RS (1993) Extended BNF—a generic base standard. Proceedings of the 1993 Software Engineering standards Symposium (SESS'93), Brighton, UK
39. Sengar V, Joshi T, Joy J, Prakash S, Toyama K (2007) Robust location search from text queries. Proceedings of th 15th International Conference on Advances in Geographic Information Systems (ACM GIS 2007), Seattle, Washington, USA
40. Silva MJ, Martins B, Chaves M, Cardoso N, Afonso AP (2006) Adding geographic scopes to web resources. Comput Environ Urban Syst 30:378–399
41. Smith J, Smith D (1977) Database abstractions: aggregation and generalization. ACM Trans Database Syst 2(2):105–133
42. Souza LA, Davis CA Jr, Borges KAV, Delboni TM, Laender AHF (2005) The role of gazetteers in geographic knowledge discovery on the web. 3rd Latin American Web Congress, Buenos Aires, Argentina, pp 157–165
43. Spaccapietra S, Cullot N, Parent C, Vangenot C (2004) On spatial ontologies. VI Brazilian Symposium on GeoInformatics (GeoInfo 2004), Campos do Jordão (SP), Brazil:CD-ROM
44. Sui DT (2008) The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. Comput Environ Urban Syst 32(1):1–5
45. Sun G, Chen J, Guo W, Ray Liu KJ (2005) Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs. IEEE Signal Process Mag 22(4):12–23

46. Tsichritzis D, Klug AC (1978) The ANSI/X3/SPARC DBMS framework report of the study group on dabatase management systems. Inf Syst 3(3):173–191
47. U.S. Census Bureau. (2003, March 2003). "108th CD Census 2000 TIGER/Line Files Technical Documentation." Retrieved March 2009, from http://www.census.gov/geo/www/tiger/tgrcd108/tgr108cd.pdf
48. Wang C, Xie X, Wang L, Lu Y, Ma W (2005) Detecting geographic locations from web resources. Proc. of the 2nd Int'l Workshop on Geographic Information Retrieval, Bremen, Germany, pp 17–24
49. Zandbergen PA (2008) A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban Syst 32(2008):214–232
50. Zong W, Wu D, Sun A, Lim E, Goh DHG (2005) On assigning place names to geographic related web pages. Proc. of the 5th ACM/IEEE-CS Joint Conf. on Digital Libraries, Denver, Colorado, USA, pp 354–362

Karla A. V. Borges holds a BS degree in Civil Engineering from the Pontifical Catholic University of Minas Gerais (1982), a MSc degree in Public Administration from João Pinheiro Foundation (1997), and a PhD in Computer Science from the Federal University of Minas Gerais (2006). She is currently a manager at PRODABEL, the information technology for the City of Belo Horizonte. Her main research interests include geographic data modeling, geographic databases, ontologies and geographic information retrieval.

Clodoveu Augusto Davis Junior received his B.S. degree in Civil Engineering in 1985 from the Federal University of Minas Gerais (UFMG), Brazil. He obtained M.Sc. and Ph.D. degrees in Computer Science, also from UFMG, in 1992 and 2000, respectively. He led the team that conducted the implementation of GIS technology in the city of Belo Horizonte, Brazil, and coordinated several geographic application development efforts. Currently, he is a professor and researcher at the Federal University of Minas Gerais. His main research interests include spatial data infrastructures, geographic databases, urban GIS, spatial data infrastructures, and multiple representations in GIS.



Alberto H. F. Laender holds a BS degree in Electrical Engineering and an MSc degree in Computer Science, both from the Federal University of Minas Gerais, Brazil, and a PhD degree in Computing from the University of East Anglia, UK. He joined the Computer Science Department at the Federal University of Minas Gerais in 1975, where he is currently a Full Professor and the head of the Database Research Group. In 1997, he was a Visiting Scientist at HP Labs in Palo Alto, California. He has served on the advisory committee of several Brazilian research funding agencies and is currently a member of ACM SIGMOD's Advisory Board and PhD Dissertation Award Committee. Prof. Laender has also served as a program committee co-chair, as well as a program committee member, for several national and international conferences on databases, digital libraries and Web-related topics. He is the author of more than 100 refereed journal and conference papers, and was one of the co-founders of Akwan Information Technologies, a Brazilian search technology company acquired by Google Inc. in 2005. Prof. Laender's research interests include conceptual modeling and database design, web data management, web information systems, and digital libraries.

Claudia Bauzer Medeiros is a full professor of computer science at the Universidade Estadual de Campinas (UNICAMP), Brazil. She is the head of the database research group in this university, and her projects center on design and development of scientific database applications, with emphasis on geographic data. She holds a PhD in Computer Science from the University of Waterloo, Canada (1985), a MSc in Informatics from PUC-Rio (Brazil) and a degree in Electrical Engineering from the same university. She is an author or co-author of about 50 papers on databases and software engineering methodologies, and has been (co)PI in over 30 research and development projects—some of which included partners in Germany, France, Argentina, Chile, and the USA. She is a member of the editorial board of the VLDB Journal and GeoInformatica.