J Braz Comput Soc (2011) 17:267–279
DOI 10.1007/s13173-011-0044-4

ORIGINAL PAPER

An ontological gazetteer and its application for place name disambiguation in text

Ivre Marjorie R. Machado · Rafael Odon de Alencar · Roberto de Oliveira Campos Jr. · Clodoveu A. Davis Jr.

Received: 5 March 2011 / Accepted: 20 September 2011 / Published online: 14 October 2011

Abstract The volume of spatial information on the Web grows daily, both in the form of online maps and as references to places embedded in documents and pages. Considering the spatial information needs of users, it is often necessary to recognize, within a document's text, the places to which it refers. This article presents a next-generation gazetteer, a toponymic dictionary which expands from the traditional cataloguing of place names and includes geographic elements such as spatial relationships, concepts and terms related to places. As such, we call it an OntoGazetteer, i.e., a gazetteer which also records semantic connections among places. The ontological gazetteer provides factual and semantic support to solving several common problems in geographic information retrieval. This paper presents the OntoGazetteer and demonstrates its applicability to a place name disambiguation problem. Along with other problem solutions to which the OntoGazetteer can contribute, we present a case study on recognizing and disambiguating place names within news sources.

Keywords Geographic information retrieval · Gazetteer · Spatial ontologies · Place name disambiguation

1 Introduction

The volume of information currently available on the Web is very large, and grows daily. Retrieving such information requires systems that can understand the needs of users, locate relevant documents, and present such documents under a relevance ranking. This is the task associated to information retrieval systems, which also deal with issues regarding indexing and storage of documents.

Users manifest their information retrieval needs in many ways, but mostly in the form of sets of keywords submitted to a search engine. Previous work [6, 12, 33, 42] has shown that a significant portion of the queries involve terms or expressions with spatial meaning, including place names and natural language expressions that denote positioning. However, getting significant results out of such queries is often difficult, because geographically relevant keywords sometimes are not understood as such by information retrieval systems. Geographic information retrieval techniques have important limitations in the recognition of spatial references and in dealing with ambiguous names (e.g., 'São Paulo' can be a Brazilian state, a city, or a soccer team). There are also difficulties in the retrieval of information constrained to a

A previous version of this paper appeared at GEOINFO 2010, the Brazilian Symposium on Geoinformatics.

I.M.R. Machado (✉) · R.O. de Alencar · C.A. Davis Jr.
Departamento de Ciência da Computação,
Universidade Federal de Minas Gerais (UFMG),
Av. Antônio Carlos 6627, ICEx, sl. 4010, CEP: 31270-010,
Belo Horizonte, MG, Brasil
e-mail: ivre@dcc.ufmg.br

R.O. de Alencar
e-mail: odon.rafael@gmail.com

C.A. Davis Jr.
e-mail: clodoveu@dcc.ufmg.br

R.O. de Alencar · R.O. Campos Jr.
Serviço Federal de Processamento de Dados (SERPRO),
Av. José Cândido da Silveira 1200, CEP: 31035-536,
Belo Horizonte, MG, Brasil

R.O. Campos Jr.
Programa de Pós-graduação em Engenharia Elétrica,
Pontifícia Universidade Católica de Minas Gerais (PUC Minas),
Av. Dom José Gaspar, 500, Prédio 03, sl. 218, CEP: 30535-610,
Belo Horizonte, MG, Brasil
e-mail: roberto.oliveira@ieee.org

geographic context. If a document can be associated to a set of places, it is possible to adjust its position in a ranking, or to filter out documents that refer to undesired locations. Recognizing a term as a possible reference to a place is usually done with the help of a *gazetteer*, a dictionary of place names [17].

Currently, there are gazetteers available on the Web, but they are based on very simple data structures, with just three components: the name of the place, its type (as defined in a feature type hierarchy), and its footprint (usually a pair of coordinates indicating its location). This concise structure leads to several limitations, which represent shortcomings in regard to the potential use of gazetteers in geographic information retrieval problems. Because the footprint is geometrically simple, spatial relationships between places cannot be used. Since semantic connections are not recorded either, gazetteers cannot resolve approximate or imprecise locations, such as "southern California", and cannot help in identifying semantic connections between place names that appear close together in text. Gazetteer contents are also an issue. Since gazetteers are notoriously hard to maintain and to expand, their coverage is usually irregular: although some include urban details on U.S. or European cities, Brazilian places are not as well covered. Some of these difficulties can be overcome by using geocoding services such as the ones available in the Google Maps API,[1] which do not make the gazetteer entries explicit, but are able to supply a pair of coordinates that corresponds to a textual description.

Regardless of such limitations (discussed in greater detail in Sect. 3), several Web-based geographic applications use information from gazetteers, as demonstrated by Goodchild and Hill [14]. We believe that gazetteers, as sources of organized information on places, can decisively contribute with the solution of geographic information retrieval problems. Therefore, this paper presents a novel conceptual schema for an enhanced gazetteer, in which the semantic connections among places can be recorded along with the usual topological connections, in order to support geographic information retrieval tasks. Such an ontological gazetteer, or *OntoGazetteer*, as proposed here, can go beyond the recognition of geographic names, allowing a more complete view of each place's semantic significance, expressed using its connections to other places and to terms and expressions that characterize it. Using this enhanced structure, we expect to support research initiatives toward solving problems such as place name disambiguation, geographic text classification, and geographic context recognition [2, 28, 34, 42]. The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 describes the conceptual database schema used to create the ontological gazetteer. Section 4 presents strategies for the application of the proposed gazetteer in the most important geographic information retrieval problems. Section 5 presents a case study in one of such problems, namely place name disambiguation. Finally, Sect. 6 presents our conclusions and a (rather extensive) list of future work.

2 Related work

Hill [17] presents the basic elements of digital gazetteers: the place name (toponym), its type, and a footprint, which indicates its location. Such components are typical of conventional gazetteers (i.e., the toponymical dictionaries usually found in atlases), and have been used as the basis for the development of the Alexandria Digital Library (ADL) Gazetteer[2] and others. Since ADL's pioneering initiative, such basic structure has also been used in other Web-based gazetteer projects.

Some works [31, 39, 40] present proposals for populating and automatically maintaining gazetteers. These works extract data from the Wikipedia,[3] which is a large knowledge base in different languages. Gouvêa et al. [15] propose a strategy for the identification of entities found in news texts, to be used in the development and updating of gazetteers. Alencar et al. [3] describe a strategy for classifying text into geographic categories through data extraction from Wikipedia to find evidence of place names in texts.

The need for semantic relationships between places in gazetteers suggests exploring ontology creation techniques. Some works have sought precisely that, considering the geospatial domain. Lopez-Pellicer et al. [24] have Geo-Net-PT 02, a geographic ontology of Portugal, an evolution of Geo-Net-PT 01 [32]. This ontology has been developed using a vocabulary, called Geo-Net, proposed by the same research group. Geo-Net uses a conceptual schema to describe places, using their name, type, relationships and footprint. It uses URIs, RDF and OWL to describe, share and codify the ontology. The initial application of the ontology is the discovery of geographic characteristics based on an attribute of a place. Janowicz and Kessler [19] present the process for developing an ontology as well as modifying a thesaurus, and present an interface of a gazetteer giving examples from hydrography.

Several information retrieval tasks can be performed with the aid of gazetteers, such as named entity recognition, place name disambiguation, geotagging, document classification, and others. Amitay et al. [5] present Web-a-Where, a system that identifies geotags for Web pages with the support

[1] http://code.google.com/apis/maps/documentation/geocoding/.

[2] http://www.alexandria.ucsb.edu/gazetteer/.

[3] http://www.wikipedia.org.

of a gazetteer. Souza et al. [36, 37] describe Locus, a geographic locator built around a gazetteer and based on a previously created ontology, OnLocus [8, 9]. Overell and Rüger [28] describe a model based on co-occurrence to solve the place name ambiguity problem, which uses a combination of heuristics and gazetteers.

As compared to these works, we propose changes in the usual gazetteer structure, and demonstrate that the enhancements can be useful for Geographic Information Retrieval (GIR) problems with an example on place name disambiguation. For this study, we have used several sources to populate the gazetteer, including official geographic data from mapping agencies, local administrations and utility companies, and also data on place-related terms and expressions from Wikipedia [3, 4]. The resulting structure is an ontological construct, which provides semantics for understanding references to places and conduct inferences based on recorded entities, as presented in the next section.

3 Ontological gazetteer

As previously mentioned, a gazetteer is a geospatial dictionary of place names, also known as a toponymical dictionary. Current digital versions are analogous to the toponymical indices usually found in printed atlases. While in an atlas each place name is associated to a generic type, a map number and a grid cell, in digital gazetteers a pair of geographic coordinates (lat-long) is used as a footprint. They can also include known variations of each place name, such as abbreviations and popular names, as well as language-specific versions.

The place's type comes from a previously compiled hierarchy, which varies among gazetteers. For instance, the Alexandria Digital Library Gazetteer (ADL) [17] has a top-level definition of feature types that includes administrative areas, hydrographic features, land parcels, manmade features, physiographic features, and regions. These in turn get more specialized, up to three more levels. On the other hand, the GeoNames[4] gazetteer defines feature codes, with the first level consisting of nine classes, with a single level of further specialization. Other digital gazetteers includeTGN[5] (Getty Thesaurus of Geographic Names) and GKB[6] (Global Knowledge Base).

Previous works [8, 9, 13, 36] point out some of the limitations of current online gazetteers, seen here as possible support tools for geographic information retrieval. The main limitations are (1) the limited spatial representation (a point or a rectangle) and absence of support for spatial relationships, (2) the absence of support for semantically complete, but geographically imprecise locations, such as 'south of France' or 'upstate New York', (3) the lack of intra-urban detail, including places often mentioned in natural language text and possibly known by non-residents, such as monuments or tourist attractions. Furthermore, the level of detail available in Web-based gazetteers seems to be lower in developing countries, such as Brazil [15].

In view of these limitations, Silva et al. [36] developed Locus, a geographic locator that uses a gazetteer as its main component. Results obtained from designing Locus suggested the creation of an ontology of places, named OnLocus, which was conducted by our group later [8, 9]. OnLocus describes spatial and semantic relationships between locations, distinguishing between the actual place and its name, a place descriptor. However, OnLocus was designed as part of an effort to extract geographic knowledge from Web pages, so it focused on indirect references to places, such as postal codes and telephone area codes [9]. In turn, the good performance of such indirect references in geographic information retrieval tasks [8] suggested that a gazetteer might be much more helpful if it could record the various types of relationship that exist between places, going beyond the topology of geographic objects and allowing the inclusion of other types of semantic relationship. In order to implement these kinds of semantically richer relationship, the gazetteer's design needs to include the flexible structure often found in ontology creation tools, such as Protégé,[7] thus becoming what we call an *ontological gazetteer*, or *OntoGazetteer*.

The word ontology comes from the Greek *ontos* (to be) + *logos* (word). In philosophy, this term is used to describe entities, events, processes and relations that exist among real world elements [10]. Considering the application of such structured knowledge sources to Web 2.0, Gruber [16] defines ontology as a 'formal specification of a shared conceptualization'. For most computing applications, an ontology offers a representation of knowledge from a certain domain, materialized to allow machines to work with the semantic content of information elements such as Web pages and natural language texts [10].

Several ontology creation languages have been proposed in the last decade. Such languages intend to represent knowledge through the definition and association of three basic elements: classes, relationships between classes, and class attributes. The most important languages have been standardized by the World Wide Web Consortium (W3C).[8] Currently, the most used language is OWL (Web Ontology Language) [19], which is an extension of the RDF (Resource Description Framework) language.

[4] http://www.geonames.org.

[5] http://www.getty.edu/research/conducting_research/vocabularies/tgn.

[6] http://xldb.fc.ul.pt/wiki/Grease.

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

[8] OWL Web Ontology Language Reference—http://www.w3.org/TR/owl-ref/.

A thoroughly built ontology can function as a valuable resource for information retrieval, since it provides a set of semantically correlated concepts, expressed as terms. For that reason, ontologies have been used in name disambiguation, document classification, query expansion and many other information retrieval tasks [13, 30, 41]. Furthermore, ontologies allow the inference of relationships between data, from which new knowledge can be achieved [1].

W3C standards propose that ontologies describe classes (or concepts) in the various knowledge domains, along with relationships among classes and their properties (attributes). Maedche and Staab [27] define the elements of an ontology O as $O : \{C, R, H^C, rel, A\circ\}$, where C is a set of concepts, R is the set of relationships, H^C is a direct relationship, *rel* is a function that relates concepts in a non-taxonomical way and $A\circ$ is a set of axioms expressed in an appropriate logical language.

Languages used to define ontologies have limitations for the representation of some domains. For instance, the creation of ontologies in the geographic domain is not simple, since OWL does not allow the direct representation of spatial information, and it also lacks specific rules to deduce spatial relationships and spatial integrity constraints [1]. Considering these limitations, Jones et al. [20] implement an ontology based on gazetteers and thesauri. Abdelmoty, Smart et al. [1] propose the use of two frameworks, plus OWL, to build and maintain ontologies of places. Huang e Deng [18] propose the use of SWRL to Express spatio-temporal rules and to build geo-ontologies. GeoNet, created by Lopez-Pellicer, Silva et al. [24] is implemented in RDF and comprises a set of classes and their properties, specified using WKT (Well-Known Text), GML (Geography Markup Language), and other geospatial standards.

In order to fulfill our proposal for a novel gazetteer which uses elements from ontology design, we established a correspondence between the usual elements of an ontology and the necessary gazetteer contents. From this correspondence, we derived a set of requirements for the structure of the ontological gazetteer, going well beyond what is usually included. Our intention is to enable researchers to use the gazetteer as a tool for situations in which there is the need to recognize the geospatial context of documents. The resulting ontological gazetteer, or OntoGazetteer, provides resources that allow it to be an important factor for problems such as detection of geographic references, place name disambiguation, interpretation of vague place names, spatial and textual indexing, and geographic relevance ranking.

We start by establishing an analogy between Concept (ontology) and Place (gazetteer). A concept can have several different names (terms in natural language); likewise, a place can have various names, with cultural, historical and language-related variations. Relationships between concepts in an ontology also correspond to spatial and semantic connections between places. Synonyms in an ontology are analogous to alternative names for a given place in the gazetteer. Hyponymy and hyperonymy are directly related to hierarchically defined spatial subdivisions, such as the ones which exist between a territory (e.g. country) and its subdivisions (e.g., states). The same applies to meronymy (a 'whole-parts' relationship, which occurs in situations such as the subdivision of a city block into land parcels). Other types of association between terms in an ontology can be mapped to spatial relationships (e.g. vicinity or proximity) or semantic relationships (e.g., the kind of relationship which exists between all cities that are state capitals in a country). We also included relationships between places and natural-language terms, both to denote ambiguity (places that have the same name as people or things) and to record associations (terms that are intrinsically associated to places). Table 1 summarizes the analogies between ontologies and gazetteers, and illustrates them with examples. As Table 1 shows, semantic relationships between places can be quite varied. Therefore, a mechanism to allow for the flexible creation and maintenance of semantic relationship types is required for the gazetteer.

Having determined a set of requirements, we now proceed to the design of a spatial database to support the implementation of the OntoGazetteer. In order to accomplish its enlarged role, the OntoGazetteer must be able to record various types of relationship between places, including spatial (proximity), topological (adjacency, containment), hierarchical (territorial subdivisions) and semantical, also recording the motivation behind each relationship. It should be possible to infer relationships between places, using the semantic properties of existing relationships. We also propose to expand the spatial representation of each place to a complete geometry, so that spatial and topological relationships can be established as needed, or recalculated as a result of data maintenance. In ontology engineering terms, places are treated as concepts. The OntoGazetteer records alternative names to a place as synonyms, also adapting the notions of hyponymy and hyperonymy to record hierarchies of territorial subdivisions.

Another proposed enhancement for the OntoGazetteer is the association of natural-language terms and expressions to each place, as related concepts. The idea is to improve the available information resources for performing typical geographic information retrieval activities, such as disambiguation and geographic context recognition. An experimental procedure for obtaining these terms has been presented by Alencar et al. [3], along with a classification procedure.

Figure 1 presents the OMT-G [7] schema proposed for the ontological gazetteer. The schema represents place names as attributes of the `Place` class. The `Alternative_Place` class maintains alternative names, abbreviations, acronyms, popular names and other variations. Each

Table 1 The correspondence between ontology structures and elements and the required contents of the OntoGazetteer

Ontology component	Definition	Equivalent in the gazetteer	Example
Concept	The meaning of a term	Place	*Belo Horizonte* (a term that is related to a certain place on Earth)
Synonymy relationship	Two terms X and Y have approximately the same meaning	Alternative names, nicknames, historical names, abbreviations and acronyms, language variations, spelling variations	*Belo Horizonte*: BH, Belô, B. Horizonte, Cidade de Minas *Lisbon*: Lisboa (Portuguese), Lisabona (Romanian), Lissabon (Dutch)
Homonymy relationship	Term X is spelled exactly as term Y, but meanings are different	Ambiguous names: places (of the same type or of different types) with coinciding names, places whose names coincide with other things.	*São Paulo* (city) and *São Paulo* (soccer team); cities of *Viçosa* (MG) and *Viçosa* (AL); city of *Vitória* (ES) and the Portuguese Word *vitória*
Hyperonymy relationship	Term X has a wider meaning than term Y	Place in a higher position in a territorial hierarchy	*Brazil* and its federal states
Hyponymy relationship	Term X has a narrower meaning than term Y	Place in a lower position in a territorial hierarchy	*Copacabana* neighborhood within the city of *Rio de Janeiro*
Association relationship	Term X is associated to term Y, i.e., there is a semantic relationship between them	Semantically related places	Cities along a road; municipalities that produce soybeans; historical sites; mining areas for iron ore

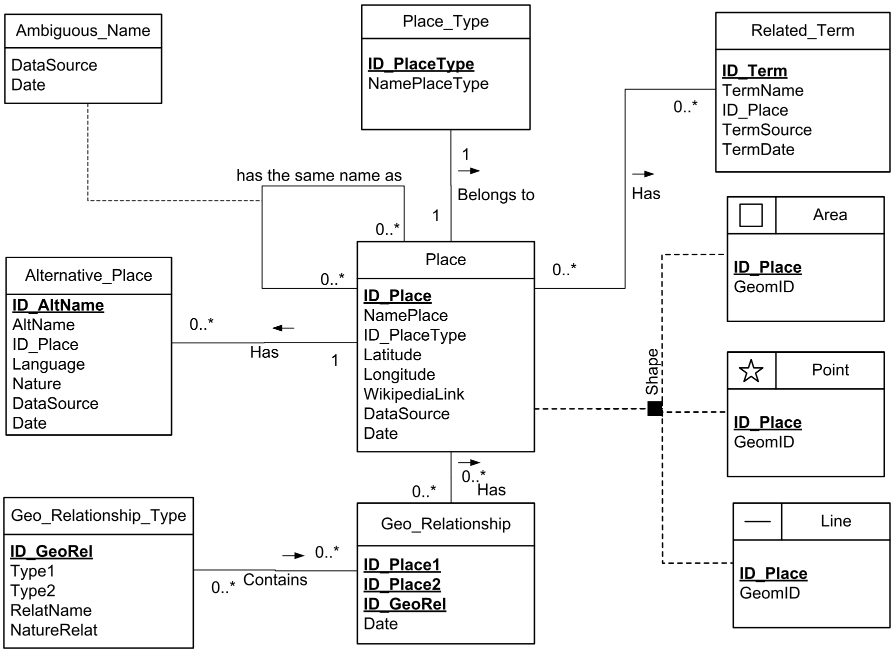


Fig. 1 Gazetteer conceptual schema

place belongs to a `Place_Type`; we initially based our place type definitions on the feature code thesaurus from ADL. Relationships between places are maintained by the `Geo_Relationship` class. Notice that two places can have several relationships between them, each one of a different `Geo_Relationship_Type`. This feature allows the gazetteer to record and use a number of different geographic and semantic connections between places. A reflexive association relationship in the `Place` class enables the creation of groups of places with ambiguous names, an information that can be helpful for disambiguation algorithms. Also for disambiguation and to support other geographic information retrieval applications, there is a class that stores lists of terms related to the place (`Related_Term`). One possible source for such names is the Wikipedia [3] (that is also the reason for keeping an attribute in the `Place` class to store the URL of its Wikipedia entry). For instance, the `Related_Term` class can contain the term 'acarajÈ' (a regional food specialty) associated to the place 'Bahia'. Finally, each place can have one or more than one geographic representation, as a point, a line or a polygon [11]. Places represented by more complex geometries will also have a point representation, as in current gazetteers.

The class diagram in Fig. 1 was detailed and mapped to a geographic database. From the OnLocus ontology, we derived several types of geographic relationship between places. Special procedures and triggers have been created for each of these relationship types, so that the relationships could be materialized in the `Geo_Relationship` table and kept up-to-date whenever new places are added. Next section describes how the features of the ontological gazetteer can be used to fulfill geographic information retrieval tasks.

4 Applications for the ontological Gazetteer

Information Retrieval (IR) has been the focus of much recent research, due to the explosive growth of the Web. Geographic Information Retrieval (GIR) expands and focuses IR techniques on problems such as the detection of references to places, or to the association of locations to Web documents. Some of these problems have been highlighted by Jones and Purves [21] in a research agenda for GIR:

1. Detection of geographic references in the form of place names;
2. Disambiguation of place names;
3. Geographic interpretation of vague place names, such as 'south of France';
4. Document indexing according to the geographic context and non-spatial content;
5. Geographic relevance ranking of documents;
6. Search interface improvement;
7. Evaluation of methods for comparing GIR systems and techniques.

Gazetteers can be used as components of the solution for most of these problems. We argue that our proposed OntoGazetteer can provide a better support for solving these and other GIR problems, since it goes beyond a simple georeferenced list of place names and introduces richer geographic and semantic relationships, related terms, and a record of ambiguous place names. In the following subsections, we will describe more specifically how the OntoGazetteer can contribute in many different GIR problems.

4.1 Detection of geographic references

Geoparsing is the process of analyzing a text in order to identify references to places, in the form of place names and other space-related terms [21]. Geotagging, on the other hand, is the process of identifying geographic entities mentioned directly or indirectly in the text and creating tags that allow the document to be linked to a location or set of locations [5, 38]. Both geoparsing and geotagging require the recognition of geographic references found intext; if this task is fulfilled adequately, the geographic context of the document can be established.

The OntoGazetteer maintains lists of official, alternative (previous names, popular nicknames) and abbreviations of place names that facilitate the identification of candidate names contained in the text. Distinguishing between actual references to places and other uses of the same words can be done by determining spatial relationships among candidate names. Since the OntoGazetteer also maintains information on spatial hierarchies and adjacent places, it is possible to infer, from the co-occurrence of related places, which candidate names should be disconsidered. Furthermore, the actual context of the document can reside in some higher level of the spatial hierarchy, e.g., a text that mentions several cities in a state actually refers to the state itself.

Notice that the proposed structure of the OntoGazetteer, in which relationships are materialized beforehand, was conceived as such in an effort to expedite relationship queries, by avoiding the execution of spatial operations during a GIR-related process. Anyway, recomputing the entire set of spatial relationships is not a major problem, since geometries are stored in the database. Applications can decide on the types of relationship that are to be considered, and which entities are to be taken into consideration, by filtering out unwanted relationships based on the relationship type. Since the full geometric shape is available, more complete and refined analyses can be performed, either in specific cases or as an additional filter.

4.2 Place name disambiguation

Place names are frequently ambiguous. For instance, 'São Paulo' exists in 6,522 different GeoNames records. According to Smith and Crane [35], 92% of TGN's toponyms are ambiguous. Several different types of ambiguity have been described in previous research [5, 41]. Amitay et al. [5] define two types of ambiguity for place names, and call them *geo/non-geo* and *geo/geo*. Geo/non-geo ambiguity occurs when a place name has also non-geographic meaning. For instance, 'Esmeraldas' can either be a city in Minas Gerais, Brazil, or the name in Portuguese of a precious stone. The same thing happens everywhere, and in every language, because places are often named after objects, people, historical facts or physical features. Geo/geo ambiguity, on the other hand, occurs when two distinct places have the same name, as in Paris (Texas) and Paris (France), which is also common, since many places are named after more famous places. We represent geo/geo ambiguity in the gazetteer using a reflexive association relationship in the `Place` class, in order to allow a more efficient retrieval of ambiguously named places.

Volz et al. [41] identified three types of ambiguity. The first type, in which the same name can refer to different places is called multiple reference. The second type corresponds to when a place is known by various names, and is called a variant name. The third type is equivalent to the geo/non-geo ambiguity defined by Amitay et al. [5], which represents places that have a name with another meaning.

Several methods are proposed in the literature to deal with ambiguity and improve the performance of geographic information retrieval systems [28]. Overell and Rüger [29] propose the application of co-occurrence models generated from Wikipedia entries to solve the place name disambiguation problem by using supervised learning techniques.

When humans read a text, ambiguities are resolved using their previous knowledge and subtle hints found in the text itself, or in elements that surround it, such as the section of a newspaper in which the text appears. Place name disambiguation, also known as toponym resolution, tries to imitate these methods [21]. The OntoGazetteer can help in this task by offering lists of ambiguously named places, alternative names and related places. These additional pieces of information can be used in heuristics designed to establish which one of the ambiguously named places is the most likely to be the one the text refers to, as we will demonstrate in the next section. The list of related terms included in the OntoGazetteer can contribute as well. If one or more of the candidate places has a weak relationship to other elements found in text (other place names, natural language terms), it can probably de disregarded.

4.3 Interpretation of vague place names

People often use vague or approximate references to places in natural language, as in ‘downtown’ or ‘Northern Italy’. In spite of the likely mention to a definite place, the geographic scope of such a reference is rough and imprecise [21]. Gazetteers usually do not include references to vague places, and the limited spatial representation keeps them from being located adequately. Using the complete geographic representation available in the OntoGazetteer, it becomes possible to infer a subdivision of the place mentioned using clues provided by the associated natural language expressions. The usefulness and interpretation of space-related expressions for GIR has been demonstrated in previous work [12].

4.4 Spatial and textual indexing

One of the techniques for indexing the contents of a text document is the creation of an inverted index file for the words contained in the document. This index provides, therefore, an association of each word to the list of documents that contain it. In the case of geographic references, this idea can be expanded using a list of places in addition to the list of words. The source for the list of places can naturally be a gazetteer [21]. After the identification of places related to each document, a spatial index can be generated, using positions (footprints) or minimum bounding boxes of the full geographic representation, so that documents can be retrieved using spatial relationships, such as proximity and containment.

4.5 Geographic relevance ranking

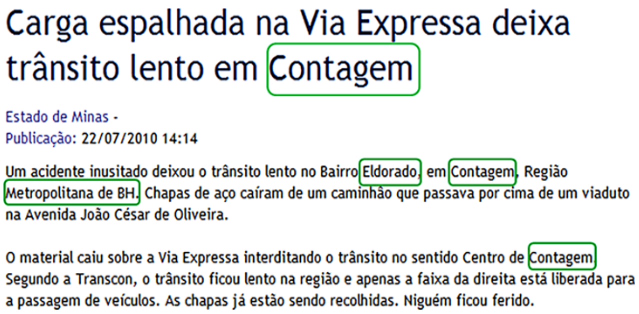
Ranking according to geographic relevance requires a measurement of the relative importance of a document for a given query. Usually, documents are selected according to the occurrence of the query terms, and ranked according to a measurement that takes into consideration the existing links to candidate documents. In the case of a geographic ranking, there must be an association of the query terms (or query region) to the places referred to by the document, and ranking needs to combine both geographic and keyword-based criteria [21]. Since the OntoGazetteer keeps lists of relevant terms, a ranking strategy can determine how specific certain query terms are in relation to places, helping to narrow down the results and assigning more importance to documents in which both terms and places are related to the query. Proximity relationship can be determined from footprints and geographic representations, so that aspect can influence the ranking as well.

5 Case study: place name disambiguation

The OntoGazetteer has been implemented in Java, using PostGIS as the spatial database management system. Currently, it includes about 150,000 places, almost 75,000 alternative names, 200,000 relationships, and 247,000 related terms. These data have been obtained from various sources, mostly official geographic data producers in Brazil and other gazetteers (for more information on the initial data sources, see [26]). An interface for interactively querying the gazetteer’s contents has also been developed.[9]

In our case study, we consider Web news sources. Usually, news texts contain one or more locations related to the facts, as part of the news reporting technique. Therefore, in this case study we put together a collection of news texts to disambiguate place names. For this, some of the strategies previously mentioned, such as detecting the occurrence of place names and inferring the geographic context of each of news texts from the implicit relationships identified from the extracted place names [25], are used.

[9] http://greenwich.lbd.dcc.ufmg.br:8080/ontogazetteer/.



Carga espalhada na Via Expressa deixa trânsito lento em Contagem

Estado de Minas -
Publicação: 22/07/2010 14:14

Um acidente inusitado deixou o trânsito lento no Bairro Eldorado, em Contagem, Região Metropolitana de BH. Chapas de aço caíram de um caminhão que passava por cima de um viaduto na Avenida João César de Oliveira.

O material caiu sobre a Via Expressa interditando o trânsito no sentido Centro de Contagem. Segundo a Transcon, o trânsito ficou lento na região e apenas a faixa da direita está liberada para a passagem de veículos. As chapas já estão sendo recolhidas. Niguém ficou ferido.

Fig. 2 Example (source: Uai—Minas—July 22, 2010)

Figure 2 presents a brief news text (in Portuguese) obtained from the UAI-Minas Web site.[10] Place names have been found in the title and in the body of the text, and are marked in the figure (*Contagem, Eldorado, Metropolitana de BH*). The presence of these names, and the spatial relationship between the places they represent, indicate that the text also refers to other places, for instance Minas Gerais state and the Belo Horizonte microregion. These places are not referred to by the text, but they can be identified from the relationships they have with the explicitly mentioned places. We say that Minas Gerais and Belo Horizonte are implicit in the text, and were identified because explicitly mentioned places relate to them.

Notice, however, that explicitly mentioned places can be ambiguous. For instance, the name *Eldorado* in Fig. 2 is associated to a neighborhood in the city of Contagem, Minas Gerais, but it is also the name of other places in the states of Espírito Santo and Mato Grosso. These versions of *Eldorado* are related to one another in the gazetteer. If we consider the other place names that occur in the text and the relationships to all versions of *Eldorado*, we can determine which version is correct, thereby achieving the disambiguation of place names in the text. However, there are many sources of ambiguity, and several situations in which it appears in natural language text. In this experimental evaluation, we present a disambiguation method that works with the support of the gazetteer and can be employed in such situations. As a baseline, we also present results from a more traditional approach, using heuristics for toponym resolution.

Resolving toponyms, i.e., determining the correct place that is to be associated to a given place name, is one of the most important problems for geographic information retrieval. Disambiguation of place names is an important part of toponym resolution. As we briefly outlined in Sect. 4.2, several methods have been proposed to deal with place name ambiguity. Usually, these methods involve using various heuristics which gradually use additional evidence or inference attempts, based on the co-occurrence of place names and on the use of other elements found in the text to reinforce one possible match over all others. Leidner [23] analyzed these alternatives and summarized the existing strategies into 16 different heuristics used to resolve toponyms in text. The application of these heuristics in sequence to a text, thus creating a disambiguation method, starts by determining non-ambiguous place names, which then serve as the basis for other heuristics that try to resolve additional names. For instance, if the text mentions two places, one of which is ambiguous, the distance from each ambiguous place alternative to the non-ambiguous one can be used as a deciding factor in favor of one version. Other heuristics decide for the most populous place, or the largest one. Naturally, the precedence of one heuristic over the others is highly dependent on the type of input and on the characteristics of the reference data, which makes it hard to compare the methods experimentally.

Leidner [22] points out that evaluation of toponym resolution and disambiguation is further complicated by the absence of a widely recognized reference dataset, which must include both a reference gazetteer and a reference corpus. To this day, no such reference dataset exists, and as a result comparative evaluations need to either build a dataset and implement all methods to be compared, or obtain datasets used in other experiments and use them with new algorithms. Furthermore, for the purposes of this paper, this second option cannot be used, since existing gazetteers lack the geographic and semantic relationship data, as well as the intra-urban detail, that distinguish the OntoGazetteer. Using just the corpus of another reference dataset was also not an option, since so far the OntoGazetteer's contents are concentrated in Brazil. We then opted for the first alternative, i.e., creating a corpus to be used along with the OntoGazetteer, in order to be able to exemplify the OntoGazetteer's use in a geographic information retrieval problem.

5.1 Creation of the news texts collection

We collected news texts from the Web between January and February 2011 (Table 2). For each of the news sources, a collector was developed in order to extract and store its title and body text, using XPath. Only news about the state of Minas Gerais was collected, because most of the gazetteer data put together so far refer to this state. In order to ensure that, news was obtained in this local- or state-related sections of the news sites.

5.2 Place name disambiguation

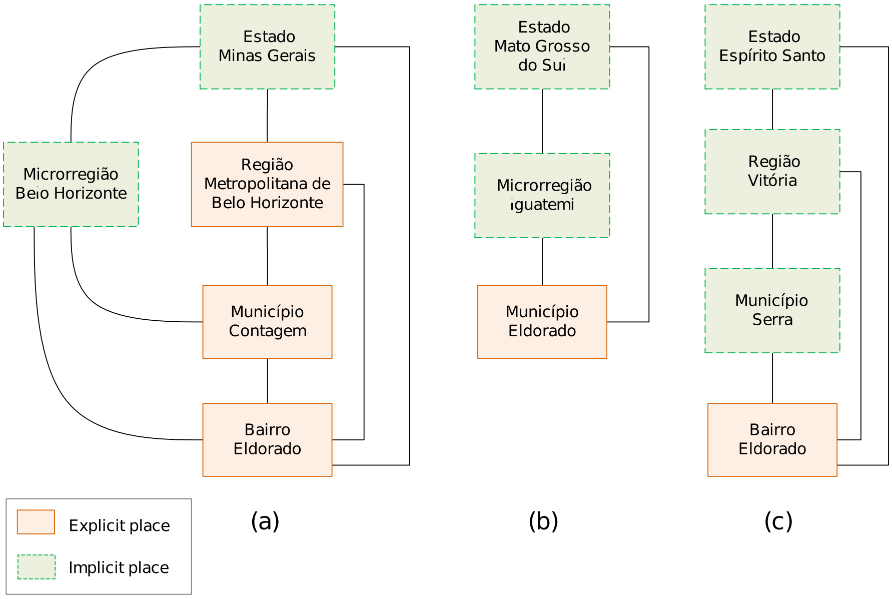
After the news documents had been collected, a pre-processing step removed stopwords (except for 'de', 'da(s)', 'do(s)', which are quite common in Brazilian place names

[10] http://www.uai.com.br/.

Table 2 Web news sources

News web	News website	Local news section name	# Docs
Globominas	http://globominas.globo.com/	General News	90
O Tempo	http://www.otempo.com.br/	Latest News (cities)	39
Uai	http://www.uai.com.br/	Minas	21
Terra	http://www.terra.com.br	Latest News (Brazil)	10
Total			160

Fig. 3 Relationship graph



and can be important for correct recognition). Next, candidate names were extracted, using regular expressions that were designed to identify single or composite proper nouns.

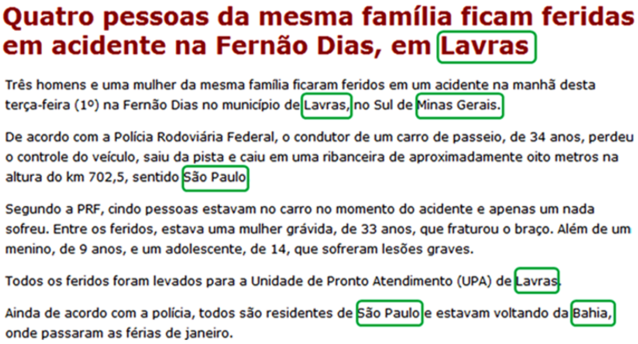
The recognition of place names from the news documents was supported by the ontological gazetteer. A simple string matching was performed between candidate names and place names from the gazetteer, including alternative names. Instances from the `Geo_Relationship` class were used to infer implicit references to other places. This inference procedure identified places whose names did not appear in the text, but were related to most of the explicitly mentioned names. Typically, names of the one that is higher in the territorial subdivision hierarchy were found [25].

Next, a first disambiguation step was executed. We begin the process by identifying non-ambiguous and ambiguous names (i.e., place names that occur only once or more than once in the gazetteer). Thus, we are considering in this study only geo/geo ambiguity. Feature type names from the gazetteer are also used in the disambiguation as *location indicators*; for instance, in Fig. 2, the word *Bairro* (neighborhood) appears before *Eldorado*, and in the gazetteer the name *Bairro Eldorado* exists as an alternative name for two different places. Of course, if there are still multiple places associated to a name after this step, disambiguation must proceed.

For the next disambiguation step, we employ the OntoGazetteer's information on relationships among places. We retrieve all recorded connections between the candidate places, both implicit and explicit, and build a graph in which places are the nodes and relationships are the links. The likely correct variation of the place name will be the one with the largest number of relationships to other places. Figure 3 shows the graph built using the place names from the text in Fig. 2. The subgraph 3(b) has been discarded by the previous step. The subgraph in Fig. 3(a) is the one which contains the most likely version of *Eldorado*, since in it this place name has connections to most of the other place names found in the text.

For a further disambiguation step, which we will not detail here, the various types of relationship recorded in the OntoGazetteer can be used. For instance, we can take into consideration only hierarchical (territorial subdivision) relationships and discard the rest, or we can also consider vicinity relationships.

The disambiguation process we presented depends on the quality and completeness of the contents of the gazetteer. In order to illustrate the effect of missing data, consider the



Quatro pessoas da mesma família ficam feridas em acidente na Fernão Dias, em Lavras

Três homens e uma mulher da mesma família ficaram feridos em um acidente na manhã desta terça-feira (1º) na Fernão Dias no município de Lavras, no Sul de Minas Gerais.

De acordo com a Polícia Rodoviária Federal, o condutor de um carro de passeio, de 34 anos, perdeu o controle do veículo, saiu da pista e caiu em uma ribanceira de aproximadamente oito metros na altura do km 702,5, sentido São Paulo

Segundo a PRF, cindo pessoas estavam no carro no momento do acidente e apenas um nada sofreu. Entre os feridos, estava uma mulher grávida, de 33 anos, que fraturou o braço. Além de um menino, de 9 anos, e um adolescente, de 14, que sofreram lesões graves.

Todos os feridos foram levados para a Unidade de Pronto Atendimento (UPA) de Lavras.

Ainda de acordo com a polícia, todos são residentes de São Paulo e estavam voltando da Bahia, onde passaram as férias de janeiro.

Fig. 4 Example (source: OTempo—February 01, 2011)

news text presented in Fig. 4. The context of this article is ambiguous, since there are references to apparently unrelated places. For instance, the place name *São Paulo* is related to a Brazilian state, a city, and a neighborhood of Belo Horizonte, Minas Gerais. The news text is ambiguous: it can either refer to the city, as the endpoint of the highway, or to the state, as the next state in the indicated direction. The connection of the neighborhood name to the reference to *Minas Gerais* can lead to misinterpretation. This problem is caused by the fact that the name *Fernão Dias*, which refers to the highway that connects Minas Gerais and São Paulo states (or, more specifically, their capital cities, Belo Horizonte and São Paulo) and runs through the city of Lavras, is the actual subject of the text. However, in this example, *Fernão Dias* is not included in the gazetteer so far, leading to a possible error. Naturally, we expect this kind of error to be less frequent as the OntoGazetteer's contents expand.

In these local or state-related news texts, we observed that omissions and implicit context are much more common than texts that talk about widely different places. Imagine, for instance, news of a sports competition involving athletes from all over the world, taking place at some city; while the text mentions many countries, perhaps its subject is the competition and the place where it is about to happen. A wider variety of places within the analyzed text could either lead to unusable results (too broad a context) or to multiple references to distinct places, with no precedence of one over the others, leading to a *ranking* problem. This aspect of the disambiguation problem definitely needs more experimentation, especially with a wider variety of textual sources, but we think the semantic connections between places in the OntoGazetteer can be helpful.

Another limitation of this technique is the fact that it is currently limited to geo/geo ambiguity. For instance, the word *Vitória* can appear in a text and be unambiguously identified as the capital city of Espírito Santo state. However, there is a soccer team with this exact name, but it is based in Salvador, the capital city of Bahia state. Depending on the occurrence or not of other place names in the text to indicate a reference to some place in Espírito Santo, a misidentification can occur. We plan to improve on that by expanding the lists of terms related to places, already included in the OntoGazetteer's structure, using Wikipedia as a source of terms [3, 4].

For the baseline, we also executed four of Leidner's [23] heuristics in a logical sequence, this time using only information that is usually available in most gazetteers. First, each place mentioned in the text was checked for ambiguity (heuristic A). The second step (B) involves comparing place names found to be ambiguous to the ones that are non-ambiguous, selecting the alternative that is closer to any of the non-ambiguous places. If this still fails to disambiguate, we then looked for a prefix which might explicitly indicate the type of place, as in 'city of Belfast' (C). Finally, if there were still ambiguous places, we selected the one that his higher in a spatial subdivision hierarchy (D). Since our OntoGazetteer includes many intra-urban place names, we considered this hierarchy to be, from top to bottom: state, mesoregion, micro region, city, neighborhood, thoroughfare. If the ambiguity still persisted, or if the disambiguated result was not the correct place, we considered that the disambiguation failed.

5.3 Experimental results

We executed the recognition and disambiguation procedures over the news texts in our collection, and manually verified the rate of success. A Web interface was created to facilitate this verification. The interface shows the text's title and body, and lists the place names identified automatically by our algorithm. For each of these place names, a volunteer would then indicate if the disambiguation worked correctly or not. Our collection initially had 160 documents, 152 of which contained place names. From those, 128 documents contained ambiguous place names. On average, each text contains four place names, of which three are ambiguous. Volunteers manually verified 100% of the disambiguated place names found in these 128 texts.

As to the baseline, we recorded results after each step, so that we could determine the most effective heuristic. We observed that many places were only disambiguated at the D step, and that non-ambiguous results in the A step were quite rare, with only 9% of the results. The weak results in the A step influenced the B step as well, since it is a heuristic based on using non-ambiguous results as a deciding factor. The use of preceding descriptors (C step) was not common, even though its use might seem obviously important. Other heuristics might have been employed, such as disambiguating in favor of the more populous place, but that would require additional information that not all gazetteers have. Results show that 346 place names were analyzed, and 82% were found to be correctly disambiguated using our proposed method and the gazetteer's relationships, as opposed

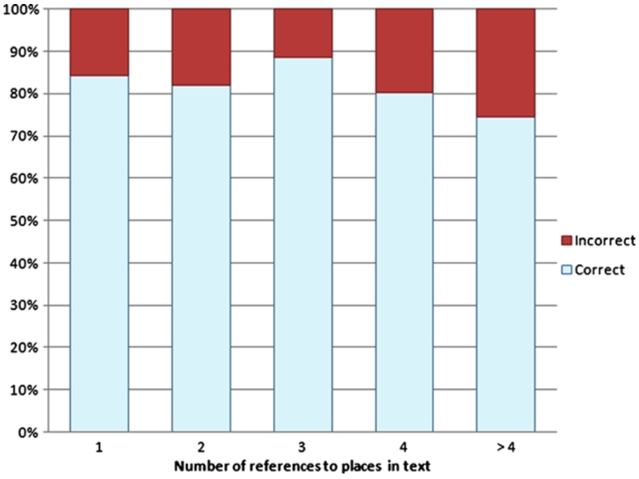


Fig. 5 Effectiveness according to the number of places mentioned in the text

Table 4 Assessment of the number of disambiguated place names

Disambiguation process	Evaluation	#	%
Using a location indicator	Incorrect	1	0.3
	Correct	63	18.2
Using relationships	Incorrect	62	17.9
	Correct	220	63.6
Total		346	100.0

Table 5 Effectiveness according to the number of places mentioned in the text

Number of references to places in text	#Correct	#Incorrect
1 place	16	3
2 place	87	19
3 place	77	10
4 place	45	11
More than 4 places	58	20
Total	283	63

to a 75% success rate using the sequence of heuristics (Table 3). The results from the intermediate heuristics show that the explicit spatial and semantic relationships available to the OntoGazetteer were more apt to determine the connections between place names mentioned in text, and therefore more effective for disambiguation.

As mentioned before, in our method disambiguation runs in two consecutive steps, one using the name of the feature type, if available, and the other building and analyzing the relationships graph. Table 4 shows the proportion of correct and incorrect disambiguations for each step. Notice that the first process was able to solve less than 20% of the ambiguities, but it was very effective, with only one incorrect result. Uncertainty is much greater in the second step, but as we demonstrated it can be dependent on the quality and extent of the gazetteer's contents.

A manual inspection of the incorrect results in the second disambiguation step showed several cases of geo/non-geo ambiguity (for instance, the word "união" is both a noun in Portuguese and a neighborhood in Belo Horizonte). In the future, this kind of problem will be more adequately treated using the OntoGazetteer's list of place-related terms. We also noticed that volunteers failed to recognize some place names as valid, since they came from seldom used and not usually known territorial subdivisions, such as microregions and mesoregions. The names of such subdivisions are sometimes significant, but some are rather obscure for the common citizen. These two cases alone account for 80% of the failures, and the remaining errors are due to various random causes, to be expected in this kind of IR experiment.

Figure 5 and Table 5 show the percentage of correct and incorrect disambiguations according to the number of places mentioned in the text. We observe that there is a small variation in the rate of success as the number of places varies. It would be reasonable to expect that the cases in which there are too few or too many place names in the text would generate a lower rate of success. For instance, with just one ambiguous place name, in any instances there would be no information to help with the disambiguation. On the other hand, if there are too many place names, the chances for confusion in the analysis of the graph would also increase the chances for errors. However, in our experiments, this was not the case: the error rate is a little higher when there are more place names in the text, but the difference to the other situations is small. At this point, however, we do not have sufficient statistical information to prove or disprove this intuitive observation.

This experiment has shown that the proposed disambiguation process is promising, as demonstrated by the results above. Notice that the quality of the results can improve in the future, since the OntoGazetteer's contents are still being expanded. We also think that this technique can improve by using other resources included in the gazetteer, such as lists of terms related to places (`Related_Term` table), to cover at least geo/non-geo ambiguity. Further study is necessary to establish the impact of the number of recognized place names in the precision of the result. These resources are currently under development, and we intend to experiment on a combination of techniques in the near future.

6 Conclusions and future work

This paper proposed a new structure for gazetteers that seeks to diminish their limitations as components of geographic

Table 3 Assessment of the number of disambiguated place names

	OntoGazetteer	Heuristic A	Heuristics A-B	Heuristics A-C	Heuristics A-D
Correct	81.9%	9.1%	26.3%	33.6%	75.0%
Incorrect or still ambiguous	18.1%	90.9%	73.7%	66.4%	25.0%

information retrieval systems. Our proposal uses ontology concepts to define a flexible way to establish and maintain semantically richer relationships between places, and adds resources for keeping alternative names and lists of place-related terms. Relationships go beyond the geographic or topologic ones, and can be used to create semantic connections between geographically unrelated places. The paper also described ways in which the semantically enhanced gazetteer can be used in typical geographic information retrieval tasks.

Naturally, the usefulness of the OntoGazetteer is a direct function of the quality and comprehensiveness of its contents. Therefore, our first task in the near future is to expand the gazetteer's contents as much as possible, using information already available in geographic databases. From geographic features found in databases, we can easily derive geographic and topologic relationships. Semantic relationships are being expanded initially considering indirect geographic relationships; e.g., two municipalities through which the same river runs are considered to be related, even though they are not adjacent to each other. Place-related terms are the focus of some parallel work in our group, using the Wikipedia as a knowledge base with promising results [3].

A case study implemented a GIR task, namely the disambiguation of references to places in news documents, and showed that the OntoGazetteer can be a valuable resource for solving that problem. Furthermore, using relationships recorded in the OntoGazetteer, we were able to infer the connection between many documents and places that are not explicitly mentioned in their text. Disambiguation achieved good results, but we feel there is room for improvement as the contents of the OntoGazetteer expand. For instance, information on features such as hydrographic basins and highways can serve as the source of relationships among different groups of places: two non-neighboring cities can be related to each other because they are located along the path of a major highway, or along the course of a river.

Future work includes developing more case studies with a broader base of documents. There is also the need to evaluate the potential use of the various relationship types and their impact in disambiguation performance. We also intend to develop a service-based interface to the gazetteer, so that remote applications can retrieve data and execute queries, without direct access to the gazetteer's database. Finally, the expansion of the gazetteer's contents, including related term lists, is our hardest but more important goal. For that effect, the use of collaborative sources, such as Wikimapia[11] and OpenStreetMap[12] is being considered, along with methods to establish data quality and filter out inadequate contributions.

Acknowledgements This work was partially supported by the Brazilian research funding agencies CNPq and FAPEMIG, and has been developed under the scope of the National Institute of Science and Technology for the Web (InWeb), also funded by CNPq.

References

1. Abdelmoty AI, Smart P, Jones CB (2007) Building place ontologies for the semantic web: issues and approaches. In: Proceedings of the 4th ACM workshop on geographical information retrieval, GIR'07. ACM, New York, pp 7–12
2. Adriani M, Paramita ML (2007) Identifying location in Indonesian documents for geographic information retrieval. In: Proceedings of the 4th ACM workshop on geographical information retrieval, GIR'07. ACM, New York, pp 19–24
3. Alencar RO, Davis CA Jr, Gonçalves MA (2010) Geographical classification of documents using evidence from Wikipedia. In: Proceedings of the 6th workshop on geographic information retrieval, GIR'10. ACM, New York, pp 12:1–12:8
4. Alencar RO Davis CA Jr (2011) Geotagging aided by topic detection with Wikipedia. In: Geertman S, Reinhardt W, Toppen F (eds) Advancing geoinformation science for a changing world. Lecture notes in geoinformation and cartography, vol 1. Springer, Berlin, pp 461–477
5. Amitay E, Har'El N, Sivan R, Soffer A (2004) Web-a-where: geotagging web content. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'04. ACM, New York, pp 273–280
6. Backstrom L, Kleinberg J, Kumar R, Novak J (2008) Spatial variation in search engine queries. In: Proceeding of the 17th international conference on World Wide Web, WWW'08. ACM, New York, pp 357–366
7. Borges KA, Davis CA, Laender AH (2001) Omt-g: an object-oriented data model for geographic applications. GeoInformatica 5:221–260. doi:10.1023/A:1011482030093
8. Borges KAV, Davis CA Jr, Laender AHF, Medeiros CB (2011) Ontology-driven discovery of geospatial evidence in web pages. GeoInformatica, to appear (available as OnlineFirst). doi:10.1007/s10707-010-0118-z
9. Borges KAV, Laender AHF, Medeiros CB, Davis CA Jr (2007) Discovering geographic locations in web pages using urban addresses. In: Proceedings of the 4th ACM workshop on geographical information retrieval, GIR'07. ACM, New York, pp 31–36

[11] http://www.wikimapia.org/.

[12] http://www.openstreetmap.org/.

10. Breitman K (2006) Web semântica: a Internet do futuro, 1st edn. LTC Editora, Rio de Janeiro
11. Davis CA Jr, Laender AHF (1999) Multiple representations in GIS: materialization through map generalization, geometric, and spatial analysis operations. In: Proceedings of the 7th ACM international symposium on advances in geographic information systems, GIS'99. ACM, New York, pp 60–65
12. Delboni TM, Borges KA, Laender AH, Davis CA (2007) Semantic expansion of geographic web queries based on natural language positioning expressions. Trans GIS 11(3):377–397
13. Fu G, Jones CB, Abdelmoty AI (2005) Ontology-based spatial query expansion in information retrieval. In: Meersman R, Tari Z (eds) Proceedings of the OTM confederated international conferences, vol 3761. Springer, Berlin
14. Goodchild MF, Hill LL (2008) Introduction to digital gazetteer research. Int J Geogr Inf Sci 22(10):1039–1044
15. Gouvêa C, Loh S, Garcia LFF, Fonseca EB, Wendt I (2008) Discovering location indicators of toponyms from news to improve gazetteer-based geo-referencing. In: Proceedings of the simpósio brasileiro de geoinformática, GEOINFO 2008. SBC, Porto Alegre
16. Gruber T (2009) What is an ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
17. Hill LL (2000) Core elements of digital gazetteers: placenames, categories, and footprints. In: Proceedings of the 4th European conference on research and advanced technology for digital libraries, ECDL'00. Springer, London, pp 280–290
18. Huang YQ, Deng GY (2009) Research on representation of geographic spatio-temporal information and spatio-temporal reasoning rules based on geo-ontology and SWRL. In: International conference on environmental science and information application technology proceedings, vol. 3, pp 381–384
19. Janowicz K, Kessler C (2008) The role of ontology in improving gazetteer interaction. Int J Geogr Inf Sci 22:1129–1157
20. Jones CB, Alani H, Tudhope D (2001) Geographical information retrieval with ontologies of place. In: Proceedings of the international conference on spatial information theory: foundations of geographic information science, COSIT. Springer, London, pp 322–335.
21. Jones CB, Purves RS (2008) Geographical information retrieval. Int J Geogr Inf Sci 22:219–228
22. Leidner JL Towards a reference corpus for automatic toponym resolution evaluation. In: Proceedings of the geographic information retrieval (GIR) workshop held at the 27th annual international ACM SIGIR conference
23. Leidner JL (2008) Toponym resolution in text. Dissertation Com Publishers
24. Lopez-Pellicer FJ, Silva MJ, Chaves M (2010) Linkable geographic ontologies. In: Proceedings of the 6th workshop on geographic information retrieval, GIR'10, ACM, New York, pp 1:1–1:8
25. Machado IM, Alencar RO, Campos R Jr, Davis CA Jr (2010) An ontological gazetteer for geographic information retrieval. In: Proceedings of the GeoINFO. SBC, Porto Alegre
26. Machado IMR (2011) Um gazetteer ontológico para recuperação de informação geográfica. Master's thesis, Departamento de Ciência da Computação da Universidade Federal de Minas Gerais
27. Maedche A, Staab S (2001) Ontology learning for the semantic web. IEEE Intell Syst 16:72–79
28. Overell SE, Rüger S (2007) Geographic co-occurrence as a tool for GIR. In: Proceedings of the 4th ACM workshop on geographical information retrieval, GIR'07. ACM, New York, pp 71–76
29. Overell SE, Stefan R (2006) Identifying and grounding descriptions of places. In: SIGIR Workshop on GIR, Seattle, Washington, pp 2–4
30. Ping D, Yong L (2009) Building place name ontology to assist in geographic information retrieval. In: Proceedings of the 2009 international forum on computer science-technology and applications, vol 1, IFCSTA'09. IEEE Computer Society, Washington, pp 306–309
31. Popescu A, Grefenstette G, Moëllic PA (2008) Gazetiki: automatic creation of a geographical gazetteer. In: Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries, JCDL'08. ACM, New York, pp 85–93
32. Rodrigues C, Chaves M (2006) Uma representação ontológica da geografia física de Portugal
33. Sanderson M, Kohler J (2004) Analyzing geographic queries. In: Proceeding of the 2nd international workshop on geographic information retrieval, GIR'04. ACM, New York
34. Silva MJ, Martins B, Chaves M, Afonso AP, Cardoso N (2006) Adding geographic scopes to web resources. Comput Environ Urban Syst 30:378–399
35. Smith DA, 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, ECDL'01. Springer, London, pp 127–136
36. Souza LA, Davis CA Jr, Borges KAV, Delboni TM, Laender AHF (2005) The role of gazetteers in geographic knowledge discovery on the web. In: Proceedings of the third Latin American web congress. IEEE Computer Society, Washington, pp 157–158
37. Souza LA, Delboni TM, Borges KAV, Davis CA Jr, Laender AHF (2004) Locus: um localizador espacial urbano
38. Teitler BE, Lieberman MD, Panozzo D, Sankaranarayanan J, Samet H, Sperling J (2008) Newsstand: a new view on news. In: Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems, GIS'08. ACM, New York, pp 18:1–18:10
39. Toral A, Munoz R (2006) A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In: EACL 2006
40. Uryupina O (2003) Semi-supervised learning of geographical gazetteers from the Internet. In: Proceedings of the HLT-NAACL 2003 workshop on analysis of geographic references, vol 1, HLT-NAACL-GEOREF'03. Association for Computational Linguistics, Strasbourg, pp 18–25
41. Volz R, Kleb J, Mueller W (2007) Towards ontology based disambiguation of geographical identifiers. In: WWW2007, Banff, Canada
42. Wang L, Wang C, Xie X, Forman J, Lu Y, Ma WY, Li Y (2005) Detecting dominant locations from search queries. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'05. ACM, New York, pp 424–431