Computers & Geosciences 96 (2016) 23–34

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Review article

A survey on the geographic scope of textual documents

Bruno R. Monteiro [a,*], Clodoveu A. Davis Jr. [b], Fred Fonseca [c]

[a] Departamento de Computação e Sistemas, Universidade Federal de Ouro Preto, Brazil
[b] Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Brazil
[c] College of Information Sciences and Technology, Penn State University, USA

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form
21 June 2016
Accepted 25 July 2016
Available online 28 July 2016

Keywords:
Geographic scope resolution problem
Geoparsing
Reference resolution
Grounding references
Information retrieval
Geographic information retrieval

ABSTRACT

Recognizing references to places in texts is needed in many applications, such as search engines, location-based social media and document classification. In this paper we present a survey of methods and techniques for the recognition and identification of places referenced in texts. We discuss concepts and terminology, and propose a classification of the solutions given in the literature. We introduce a definition of the Geographic Scope Resolution (GSR) problem, dividing it in three steps: geoparsing, reference resolution, and grounding references. Solutions to the first two steps are organized according to the method used, and solutions to the third step are organized according to the type of output produced. We found that it is difficult to compare existing solutions directly to one another, because they often create their own benchmarking data, targeted to their own problem.

Contents

1. Introduction

The demand for geographic data in applications on the Web is increasing. One of the most important resources to support this increased interest is the ability to recognize references to places in

* Corresponding author.
E-mail addresses: bruno@decsi.ufop.br (B.R. Monteiro), clodoveu@dcc.ufmg.br (C.A. Davis Jr.), fredfonseca@ist.psu.edu (F. Fonseca).

http://dx.doi.org/10.1016/j.cageo.2016.07.017
0098-3004/

Web documents. If documents can be correctly and efficiently linked to places mentioned directly or indirectly in them, it becomes possible to improve and innovate in directions such as geographic indexing and querying, finding relationships based on spatial proximity or containment, and detecting localized trends for events and phenomena mentioned in social media.

A large share of the information available on the Web is geographically specific (Delboni et al., 2007; Vaid et al., 2005; Vasardani et al., 2013). References to geographic locations appear in the form of place names, postal addresses, postcodes, historical dates, demonyms, ethnicity, typical food and others. Many queries include place names and other geographic terms (Delboni et al., 2007; Sanderson and Kohler, 2005; Silva et al., 2006). Therefore, there is demand for mechanisms to search for documents both thematically (for instance, using a set of keywords) and geographically, based on places mentioned or referenced by the text (Zong et al., 2005). Similar techniques and resources can also apply to streaming data, such as Twitter messages or RSS feeds, providing the opportunity to index content in near-real-time, based on references to places.

However, while finding references to places in Web documents, ambiguity and uncertainty occur. Places can share a name with other places (Paris, besides being the capital of France, refers to more than sixty places around the world[1]). Places are named using common language words (Park, Hope and Independence are American cities) and proper names (Washington, Houston and San Francisco). The first type of ambiguity occurs when a place name references multiple places, and it is called *Geo/Geo ambiguity*, or *referent ambiguity*. The latter ambiguity is called *Geo/Non-Geo ambiguity* (*referent class ambiguity*), which occurs when both a location and a non-location share the same name (Amitay et al., 2004). Clough et al. (2004) suggest a third type of ambiguity, named *reference ambiguity*, which occurs when a place is associated to many names, like New York, NYC or The Big Apple. Ambiguity makes the resolution of references to places intrinsically context-based. Although there are important work on place-based information integration and retrieval, areas such as disambiguation are still in its infancy (Vasardani et al., 2013).

An important resource to address disambiguation is the determination of the *geographic scope* of the document, i.e., the set of places referenced by and relevant to the contents of the document. 'Every document has a geographical scope" (Andogah et al., 2012). Even keyword queries to search engines can have a geographic scope (Alexopoulos and Ruiz, 2012; Silva et al., 2006), since query words embed the user's intentions in the search.

References to places can be straightforward and unambiguous as geographic coordinates or not. Other sources of geographic location information can be structured (postal addresses) or unstructured (place descriptions in text). They can also be direct (place names) or indirect (references to cultural characteristics associated to places), explicit (news headers) or implicit ("9/11"). Humans are often able to recognize references to places based on such evidence, but this association does not come so easily to automated systems. Addressing this problem is one of the pressing tasks for Geographic Information Retrieval (GIR) research.

GIR extends Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999) with use of geographic locations and metadata (Jones and Purves, 2009), taking it beyond the use of keywords. GIR studies methods and techniques for the retrieval of information from unstructured or partially structured sources, including relevance ranking, based on queries that specify both theme and geographic scope (Jones and Purves, 2008, 2009). One of the most important research subjects currently in GIR involves recognizing references to places in regular text, and also in other media, such as photos and videos (Luo et al., 2011), including implicit references. The recognition of references to places in media other than text documents is beyond the scope of this paper.

Many initiatives to tackle the GIR problem of recognizing references to places in text have arisen in the recent past, usually with varied or conflicting descriptions or terminologies, targeting various applications, or using a range of reference data. The main contributions of this survey are (1) a structured and comprehensive view on contributions to the problem of determining the geographic scope of documents, (2) a review of relevant definitions and concepts, (3) a discussion of the main techniques currently used to address the Geographic Scope Resolution (GSR) problem, and (4) a proposal for a future research agenda in the area.

This paper is organized as follows. Section 2 introduces the main application areas related to the determination of the geographic scope of documents. Section 3 discusses the terminological variations found in the literature and presents them in a structured way. Section 4 focuses on the GSR problem, presenting a set of methods and a discussion of existing techniques from the literature. Finally, Section 5 shows our conclusions and indicates future research directions for the field.

2. Main application areas

In this section, we present a high level description of what can be gained if efficient methods for determining the geographic scope of documents are available. The main application areas are divided into two groups: (1) contributions to IR, in the form of tools and techniques that incorporate geographic variables, and (2) contributions to Web data mining.

2.1. Information retrieval tools and techniques

Most of the initial work related to the geographic scope of documents comes from information retrieval (IR), largely due to the interest in evolving Web search engines by including geographic features (Amitay et al., 2004; Buyukkokten et al., 1999; Ding et al., 2000; Gravano et al., 2003; McCurley, 2001; Silva et al., 2006; Vaid et al., 2005). Query expansion, filtering, ranking of results and other IR techniques were adapted to use geographic data as soon as the solutions to establish the geographic scope became more efficient and more scalable. We highlight four techniques here, associated to various stages of a search engine's pipeline: geographic indexing, query expansion, recognition and use of place names included in queries, and geographic ranking.

- *Geographic indexing*: Places can be associated to documents using a spatial/geographic index to support IR operations. Furthermore, collections of documents that refer to a place can be put together. Geographic indexing can also help in finding ideal locations for documents in a distributed storage infrastructure, considering that users from some geographic region tend to concentrate their interest on documents of local scope (Lieberman et al., 2010; Vaid et al., 2005).
- *Query expansion*: Searches performed using a place name can be expanded using names of topologically-related places (i.e., neighboring places), as well as places that belong to the same territorial subdivision hierarchy (i.e., a search for documents related to a state can include documents referring to any of its cities) (Andogah et al., 2012; Delboni et al., 2007; Machado et al., 2011; Moura and Davis, 2014).
- *Use of place names in queries*: Traditional search engines based on keyword matching did not consider that a keyword might

[1] According to GeoNames: http://www.geonames.org

represent a geographic entity, and therefore a geographic location (Cardoso, 2011; Delboni et al., 2007; Martins et al., 2007). More recently search engines have started to consider place names in queries leading to improved search results.
- *Geographic ranking*: Search engine users, particularly mobile ones, increasingly expect the search system to know their current geographic position and consider it in the ranking of results, by extrapolating on their geographic context (Alexopoulos and Ruiz, 2012; Alexopoulos et al., 2013). Metadata can be captured to establish the relevance of a document as to the perceived interest of the user. Semantics can be used to determine how close a result is to the user's expressed intentions based on the geographic nature of the term that is being sought and from the user's search history (Andogah et al., 2012).

2.2. Web data mining

In the late 1990s, Buyukkokten et al. (1999) presented the idea of determining the geographic scope of Web resources and proposed improvements in search engines to rank the results considering the geographic distance between the user's location and the places mentioned in the documents. They also suggested the use of geographic information to target product sales, expanding the uses of data mining applications. Two sets of techniques are prevalent in this area, geographic filtering and document classification.

- *Geographic filtering*: In streaming data sources, such as RSS feeds and social network messages, geographic data can provide criteria for prioritization and filtering of messages and documents according to the user's location. This is particularly interesting for news, traffic, and weather applications (Lieberman and Samet, 2012; Ribeiro et al., 2012).
- *Document classification*: Documents can be classified and grouped by their geographic scope. This classification makes it possible to create spatially-aware services such as map-based news selection applications. Besides that, data mining algorithms and methods can use the geographic scope to classify or to cluster documents and Web pages related to a given place (Alencar et al., 2010; Alencar and Davis, 2011; Morimoto et al., 2003; Teitler et al., 2008; Silva et al., 2006).

The applications listed in this section are not exhaustive, and many more can be envisaged, in conjunction with techniques from machine learning (Anastácio et al., 2009; Gravano et al., 2003), data mining (Backstrom et al., 2010; Lieberman and Samet, 2012), and natural language processing (Drymonas and Pfoser, 2010). Many end-user applications can also benefit from solutions to the geographic scope resolution problem, including, but not limited to, map-based browsing (Lim et al., 2002; McCurley, 2001; Teitler et al., 2008) and user location inference (Davis et al., 2011; Yang et al., 2011; Ribeiro et al., 2012). The next section presents the problem of determining the geographic scope of documents and basic concepts related to it.

3. The geographic scope resolution problem

In this section, we review a set of concepts and the terminology used the description of algorithms and techniques addressing the Geographic Scope Resolution (GSR) problem. Key and ground terms and their respective definitions are extracted from the relevant literature and adjusted to the proposed definition of the GSR problem, presented in more detail in Section 3.2.

3.1. GSR terminology

Toponym: A place name, i.e., a name that can be used to refer to places on Earth. The names are not necessarily unique, therefore different places can share the same name. Places are also named after people, other places, and other entities (Jones and Purves, 2008; Habib et al., 2013). Leidner (2007), Leidner (2008) presents an extended discussion on place names, including a historical perspective.

Toponym recognition: The process of identifying place names in text, also known as *toponym extraction* (Habib et al., 2013; Jones and Purves, 2008).

Toponym resolution: The process of mapping a toponym to an unambiguous set of spatial coordinates corresponding to a geographic location (Leidner, 2007, 2008). Also known as *toponym disambiguation* (Habib et al., 2013), *location normalization* (Li et al., 2002) or *geographical entity resolution* (Alexopoulos and Ruiz, 2012).

Geographic Scope of Documents or GS(D): The GS(D) is the set of places and/or regions associated to a document's content (Andogah et al., 2012; Buyukkokten et al., 1999; Ding et al., 2000). The GS (D) considers the places mentioned in the document, but not necessarily has to consider all of them. Besides that, in its most trivial form, the GS(D) can be expressed as a set of coordinates of the places mentioned in the document. However, the GS(D) might represent these places more broadly if necessary. For instance, if many states of a country are mentioned in a document, a more adequate GS(D) may indicate the country itself as the document's scope. Some other terms are equivalent: *geographic document footprint* (Fu et al., 2005; Silva et al., 2006), *geographic path* (Vargas et al., 2012b), and *geographic focus* (Amitay et al., 2004; Chen et al., 2010; Zubizarreta et al., 2008).

Gazetteer: A repository of georeferenced place names, usually enhanced with further information, such as the type/class, geographic coordinates (usually called a *footprint*), and some conceptual or territorial hierarchy (Hill, 2006; Leidner et al., 2003). In other words, gazetteers are dictionaries of toponyms (Hill, 2000, 2006). Gazetteers are instrumental as sources of valid place names in place name disambiguation, and in establishing the location of places identified by name (Souza et al., 2005).

The next section presents a definition of the geographic scope resolution problem, in which the variety of terminologies used in referenced works are considered and standardized.

3.2. GSR: a definition

The Geographic Scope Resolution (Alexopoulos and Ruiz, 2012; Andogah et al., 2012; Alexopoulos et al., 2013) problem consists in discovering places related to the contents of a textual document, disambiguating them if necessary, and using the resulting set of unique places to build the overall geographic scope. GSR is known under other names in the literature, with essentially the same definition, except for terminological differences: *Place Name Assignment Problem* (Zong et al., 2005; Amitay et al., 2004) or *GeoReferencing* (Gouvêa et al., 2008; Zubizarreta et al., 2008).

The problem of determining the geographic scope of a document can be structured in a sequence of tasks, as follows (Fig. 1):

1. *Geoparsing*: find all the references to places contained in the document;
2. *Reference resolution*: disambiguate these references, so that each can be associated to a unique place;
3. *Grounding references*: obtain a geometric description of each disambiguated place, such as a footprint or a set of coordinates, thus creating the geographic scope of the document.

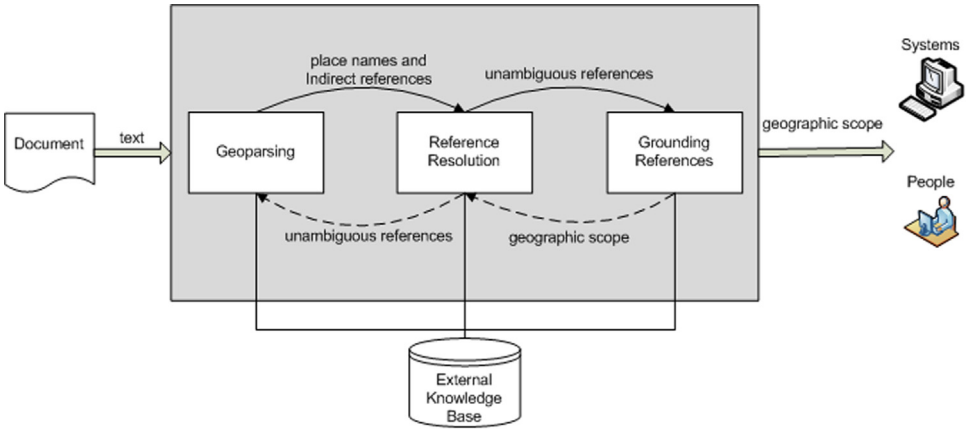

Fig. 1. The GSR problem.

Tasks (1) and (3) are mandatory. Task (2) is required in most cases, due to the existence of ambiguity (Amitay et al., 2004).

Geoparsing includes toponym recognition, but can also deal with indirect references to places in text. Likewise, reference resolution uses toponym resolution, but includes disambiguation based on evidence that goes beyond toponyms found in the text. These definitions are meant to encompass the broad range of methods we describe in more detail in the next sections.

Fig. 1 presents the GSR problem schematically. The input is represented by a document or a set of documents, and the output is the geographic scope, which can be used by a system as part of a more complex task or directly presented to users. White rectangles indicate each of the three tasks. Solid arrows indicate the direct sequence of steps to solve the problem. Dashed arrows indicate that the results of a later step can contribute to the resolution of a previous task in further iterations. The effects of disambiguation on geoparsing have already been studied and called "reinforcement effect" (Habib and van Keulen, 2011), indicating that tasks 1 (Geoparsing) and 2 (Reference resolution) can be highly dependent.

Andogah et al. (2012) paradoxically use a preliminary geographic scope, estimated using a scoring system applied on candidate toponyms, to support disambiguation, showing that the process can be cyclical.

The next sections provide a generic view on each task. Literature contributions to the global problem and to each task are presented in Section 4.

3.2.1. Task 1: geoparsing

Geoparsing is the first required task in solving the GSR problem. The objective is to find all references to places present in a document, which can be direct (toponyms) or indirect (mentions to other entities that are readily associated to a place or a geographic location).

Geoparsing can be carried out using external evidence on place names and indirect references to places. Candidate terms are searched in gazetteers or geographic databases, using string matching or special text elements such as capitalized words in sentences. Indirect references to places are called *implicit geographic evidence* (Cardoso et al., 2008) or *location indicators* (Leveling et al., 2006). They are urban addresses, references to related entities or landmarks, nicknames, or even sets of coordinates.

Machine learning algorithms and other strategies can also be used for geoparsing, which can be seen as a specialized version of Named Entity Recognition (NER) (Nadeau and Sekine, 2007). NER aims at identifying any kind of entities mentioned in natural language sentences, including places, people and objects. The most common NER algorithms used in geoparsing (Lieberman and Samet, 2011) are the Support Vector Machine (SVM), the Hidden Markov Model (HMM), and the Conditional Random Field (CRF).

Some proprietary and open-source systems can be used to find toponyms, geographic phrases and other names in text. Examples include C&C tagger,[2] Apache's OpenNLP,[3] OpenCalais,[4] Yahoo!'s Placemaker,[5] Sheffield's GATE,[6] Ling-Pipe,[7] and Stanford University's open source tagger.[8]

3.2.2. Task 2: reference resolution

Reference resolution is the process of mapping a place name or reference in documents to an unambiguous identification of the place. This definition also applies to toponym resolution.

This step is mandatory whenever the source data contain ambiguities, which is a very common problem in place names. According to Habib et al. (2013), around 46% of the toponyms found in GeoNames refer to more than one place. For instance, the toponym "Springfield" corresponds to more than one hundred and eighty different geographic places around the world, including places in the U.S., Australia, Jamaica, South Africa, and Canada.

Reference resolution usually relies on an external knowledge base, such as a gazetteer, a geographic database, or a combination of them. Toponyms and other references are used to search the knowledge base for candidate places. Additional evidence, from the document or from external sources, is then used to decide which place is more likely to correspond to the reference in the text. The quality of the results in this step is highly dependent on the quality and coverage of the external reference data.

3.2.3. Task 3: grounding references

Grounding references corresponds to the process of mapping each reference to a footprint, which may be a set of latitude and longitude coordinates, or a set of polygons, representing geographic boundaries. When multiple references to places are found in the document, this step has to deal with the granularity of the geographic scope itself. This means that some decisions should be taken considering the desired level of generalization. Consider, for instance, that a document mentions a number of neighboring cities. The result of this step can be the set of cities itself or a region containing all the mentioned cities.

[2] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Taggers
[3] https://opennlp.apache.org
[4] http://www.opencalais.com
[5] http://www.programmableweb.com/api/yahoo-placemaker
[6] https://gate.ac.uk
[7] http://alias-i.com/lingpipe
[8] http://nlp.stanford.edu/software/tagger.shtml

Other terms are used to refer to grounding references or something very similar to it. While some terms, such as *grounding* and *localization* (Amitay et al., 2004), are clear other, such as *geotagging* or *geocoding*, have different meanings in similar fields. Geotagging as a GSR term is the process which involves the identification of place names in texts and the assignment of spatial coordinates to them (Lieberman and Samet, 2011). Notice that this use is different from the most common meaning of geotagging, which is the process of creating tags that allow the document (or other types of Web objects, such as photos or videos) to be linked to a location or set of locations (Amitay et al., 2004; Teitler et al., 2008). In this case, geotags are directly added by the creator of the object, or collaboratively by other users.

Although *geocoding* is sometimes used in relation to GSR, its most common meaning is not. Other areas use *geocoding* to describe the process of locating points on the surface of the Earth based on alphanumeric address information (Davis and Fonseca, 2007) or more broadly meaning the location of places based on any kind of textual description (Goldberg et al., 2007).

4. Proposed solutions to the geographic scope resolution problem

This section presents a number of proposals from the literature for the solution of the GSR problem. The contributions are divided in two groups. First, Section 4.1 presents proposals that cover the entire GSR problem, although sometimes using concepts and steps that are slightly different from the definitions presented in Section 3. Next, Section 4.2 covers contributions that are specific to the geoparsing step. Section 4.3 presents proposals connected to the reference resolution and Section 4.4 covers proposals dealing with grounding references.

4.1. Global solutions

Some initiatives in the literature take a global approach to the GSR problem. In general the idea is to find toponyms along with other space-related terms and indirect references in order to infer the geographic scope of the documents.

Early work proposed the association of documents to the location of the Web server that hosts them (Buyukkokten et al., 1999) or to the location of the reader (Wang et al., 2005). Even though such a correspondence may exist in many cases, nothing keeps a Web server from hosting content that is unrelated to its place. Hybrid approaches tried to correct this problem using both server location and document contents (Ding et al., 2000; McCurley, 2001).

In order to use document contents, most approaches divide the work in steps that closely resemble the tasks described in Section 3.2: geoparsing, reference resolution, and grounding references. For geoparsing, there is the need to identify references to places, which can take on many forms: toponyms, urban addresses, telephone and postal area codes, and others (McCurley, 2001; Borges et al., 2011, 2007). Checking the validity of candidate names as indicators of location requires using reference datasets, such as gazetteers, ontologies or proper databases, along with matching algorithms and heuristics to take care of disambiguation. Strategies used to solve this problem are the use of a controlled dictionary as support in the geoparsing step and heuristics to resolve the ambiguity between the references (Zubizarreta et al., 2008), the use of a NER procedure and a graph-ranking algorithm, based on PageRank (Page et al., 1999) for the geoparsing and reference resolution steps (Silva et al., 2006). Zong et al. (2005) use a third party software to perform the geoparsing and a rule-based approach to disambiguate the toponyms using a tree structure to build the geographic scope instead of the graph structure used by Zubizarreta et al. (2008) and Silva et al. (2006).

Vargas et al. (2012a), Vargas et al. (2012b) and Zhang et al. (2012) also have multistage methods using a third part tool in the geoparsing step. The reference resolution step is very different in each case. Vargas et al. (2012a), Vargas et al. (2012b) use Geographic Information Systems (GIS) functions, while Zhang et al. (2012) run a disambiguation procedure, GeoRank, which adapts PageRank to solve the geo/geo ambiguity, and heuristics to solve the geo/non-geo ambiguity.

The sequence of tasks is not universally adopted, as there are works that structure the process differently, or start with a definite set of assumptions and boundary conditions. Borges et al. (2007, 2011), instead of geoparsing only for toponyms, focus on recognizing indirect reference such as urban addresses and their components: street names, telephone area codes, urban landmarks and postal addresses, using an ontology and regular expressions. In the K-Locator system (Alexopoulos et al., 2013), the definition of which concepts will be used both in the disambiguation of terms and in defining the geographic scope of the text has to be made by the user. This implies that the solutions works only on specific application scenarios: both the document's domain(s) and the nature of its contents have to be known *a priori* or must be predicted. Furthermore, comprehensive ontologies covering these domain(s) must be available. Andogah et al. (2012) use a multistage method with a different sequence of steps, first using external knowledge on toponyms to find the geographic scope, and only then executing the reference resolution step. References to places can also be indicated by sentences that contain expressions that denote positioning, such as spatial prepositions (e.g., near, inside, in front of) (Delboni et al., 2007) and lexical constructors that indicate spatial information (Woodruff and Plaunt, 1994), in approaches that border on Natural Language Processing (NLP).

An important global approach, with a particular strategy, has been proposed by Leidner (2007, 2008), who found 16 different heuristics in previous work (Amitay et al., 2004; Li et al., 2003; Olligschlaeger and Hauptmann, 1999; Pouliquen et al., 2006; Rauch et al., 2003; Smith and Crane, 2001; Smith and Mann, 2003; Woodruff and Plaunt, 1994; Zong et al., 2005). The use of these heuristics in sequence on 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.

Finally, the places that have been identified from the document compose the geographic scope. Many works that take the global approach focus on obtaining a single integrated scope from the set of identified places (Amitay et al., 2004; Campelo and Baptista, 2008; Chen et al., 2010). This is achieved using a hierarchical structure built from relationships obtained using a gazetteer or knowledge base. Amitay et al. (2004) present some heuristics to resolve references based on the “one sense per discourse” principle, which states that if an ambiguous toponym is mentioned more than once in a text, all references should correspond to the same place. Another heuristic, which consider that place names appearing in a given context tend to indicate nearby locations, is also used.

A major issue for global approaches is the lack of an established benchmark from which to compare the proposals. Each article defines its own corpus, queries, and comparison methodology. Therefore, there is currently no direct way to establish which approaches are more efficient. Anastácio et al. (2009) tried to empirically compare some proposals: the Web-a-Where system (Amitay et al., 2004); the spatial overlap-based method proposed in the GIPSY project (Woodruff and Plaunt, 1994); the graph-based method originally proposed in the GREASE project (Silva et al.,

2006); and three simple baseline methods. They concluded that the Web-a-Where system achieved the best results, closely followed by the GraphRank method and by the baseline based on the most frequently occurring place. Nevertheless, the comparison methodology cannot be used more broadly to compare methods that have been proposed more recently without being modified. However, recently, in October 2015, the International Workshop on Recent Trends in News Information Retrieval[9] released an annotated dataset that consists of 1 million news articles from a wide range of sources. This corpus can be used by researchers in the future as a benchmark dataset.

Table 1 summarizes the works that deal directly or indirectly with the GSR problem.

More recent works tend to concentrate on a single task in the GSR problem, due to the increasing complexity of the proposals. The focus on single tasks allows for more compartmentalized solutions, with sets of techniques directed at each part of the problem. Solutions for each task are discussed in the next sections.

4.2. Geoparsing solutions

This section proposes a classification of the most representative methods found in the literature for geoparsing (Fig. 2). There are three main groups of geoparsing methods: *lookup-based*, *rule-based* and *machine learning-based (supervised)* (Leidner and Lieberman, 2011). For the first two categories, methods are further classified as *language-dependent* or *language-independent*. Supervised methods are generally language-independent while most other methods are language-dependent. Some authors use consider only two categories of extraction techniques, machine learning and rule-based approaches (Habib et al., 2013), instead of the three mentioned above (Leidner and Lieberman, 2011).

In the *lookup-based* approach, each document is analyzed, and each candidate word or set of words is matched against some external data source, such as a gazetteer, an ontology or a database. The quality of the external reference data has a direct impact in the quality of the results. If the resource is incomplete, or if there is much ambiguity, the method can lead to false positives and false negatives. Methods are language dependent if the external data sources are likewise language-dependent, and multilingual reference data can be used to promote language independence.

There are several lookup-based proposals in the literature, which basically use heuristics to identify candidate strings, and then compare them to a reference dataset (Alexopoulos et al., 2013; Amitay et al., 2004; Chen et al., 2010; Clough, 2005; Pouliquen et al., 2006; Purves et al., 2007; Zubizarreta et al., 2008). There are also more complex proposals using a more varied strategy applied for obtaining candidates. Borges et al. (2007, 2011) use an ontology and regular expressions in the task of geoparsing both for toponyms and indirect references, along with a gazetteer. Similarly, Shi and Barker (2011) use a combination of gazetteer and linguistic heuristics (prepositions near the toponym and spatial relationship terms, sometimes called *spatial prepositions*) (Mark, 1989) to geoparse for elements such as provider location and domain, markup components that contain coordinates, indirect references (postal codes, phone numbers) and toponyms.

Rule-based methods use some heuristics and a set of symbolic rules, such as regular expressions or context-free grammars. Both heuristics and the rules are encoded in domain-specific language, resulting in, most cases, language-dependent methods. Woodruff and Plaunt (1994) pioneered the use of a rule-based method to geoparse texts, looking for place names near lexical constructs that

Table 1
Solutions to the GSR problem.

	Infrastructure	Infrastructure and content	Content		
			Toponym	Indirect Reference	(Both)
Early workstext	Buyukkokten et al. (1999), Ding et al. (2000), McCurley (2001), Woodruff and Plaunt (1994)				
Heuristics and look-up based methods			Amitay et al. (2004), Leidner (2007), Leidner (2008), Silva et al. (2006)		
Multistage methods		Wang et al. (2005)	Chen et al. (2010), Vargas et al. (2012a), Vargas et al. (2012b), Zhang et al. (2012),Zong et al. (2005), Zubizarreta et al. (2008)	Campelo and Baptista (2008)	Alexopoulos et al. (2013), Andogah et al. (2012), Borges et al. (2007), Borges et al. (2011)

[9] http://research.signalmedia.co/newsir16/index.html

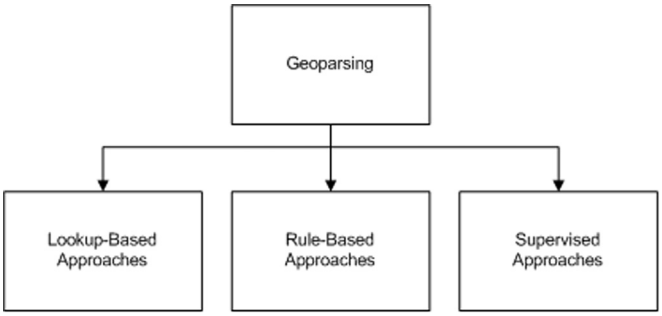

Fig. 2. Geoparsing approaches.

indicate spatial position. Silva et al. (2006) and Lieberman and Samet (2011) use a part of speech (POS) tagger to find proper nouns, since place names tend to be capitalized. These works are language-dependent, built for texts in English. Twaroch et al. (2008) also use regular expressions and filters to find place names.

On the other hand, Pouliquen et al. (2004) present a rule-based approach for geoparsing multilingual texts. First they try to discover in which language a text is written, using an n-gram-based language guesser. Then, a set of regular expressions is created to find all possible place names – countries and cities – including suffixes and other combinations. This is necessary because the same place name may have different spellings in different languages.

Machine learning-based methods are built using a training annotated corpus containing the text associated to the expected set of related places. The annotated corpus is then used to train an algorithm, using features such as infrequent strings, length, capitalization and other features. After that, the method runs on a corpus of unannotated documents and the same features are computed to decide on the association of places and documents. Techniques such as support vector machines (SVM), conditional random fields (CRF) and hidden Markov models (HMM) can be used to solve the geoparsing step.

SVM is a supervised machine learning algorithm used in data mining tasks, such as classification and regression analysis. In the SVM model, the representation of examples to be classified is done by mapping points in space, so instances of different categories can be clearly separated by a hyperplane. Then, new examples are mapped into the same space, hence their categories can be predicted (Hearst, 1998). SVM can work with two (binary SVM) or more (multi-class SVM) classes. Another example of a machine learning classifier is HMM, which is a sequence classifier. More precisely, HMM is a ubiquitous tool for representing probability distributions over sequences of observations (Fujiwara et al., 2009). HMM classifies single objects into classes, considering the characteristics of objects in the neighborhood. Similar to HMM, the CRF approach is also a supervised learning technique in which statistical models using sequence data are built. It uses an undirected graphical model that defines a log-linear distribution over sequences, given a particular sequence.

Chasin et al. (2013) present and experimentally compare three machine learning methods to find toponym candidates: (1) a SVM approach, (2) a HMM implemented by the LingPipe[10] library, and (3) the NER tool designed by the Stanford Natural Language Processing Group,[11] which uses the CRF algorithm. In order to decide whether each candidate is actually a toponym, a lookup is made using Google Geocoder.[12]

An example of a supervised method is presented by Habib et al. (2013). They use HMM and SVM to extract toponyms from a set of descriptions of holiday homes. HMM, trained using manual annotations, is used to extract candidate toponyms from the document. Candidates are then matched against data from GeoNames, generating two sets of features (positive and negative candidates) that are used to train a SVM classifier. SVM is then used as a disambiguation technique, reinforcing the results from the original extraction. The method is language-independent, as demonstrated by experiments involving texts in English, German and Dutch. In a different approach, Nissim et al. (2004) use the Curran and Clark maximum entropy tagger (Curran and Clark, 2003) to recognize location names in historical descriptions of Scotland. The tagger, as a supervised approach, is trained using 10-fold cross-validation on an annotated dataset. The method shows significant improvement in precision, recall and f-score over a lookup-based method that used a custom-built Scottish gazetteer.

There is a complicating factor in the analysis of the various methods described in this section. Results from lookup-based methods are highly dependent on the external reference source, and each proposal potentially uses a different one, including custom-built gazetteers. Rule-based approaches use resources such as regular expressions and linguistic heuristics, custom-built in many cases, with language variations and adaptations that are tuned to particular problems. Supervised methods require labeled training data, and no standard for comparing the performance of methods has emerged so far. Notice, however, that Rule-based and supervised methods also use external data, thus their results is also dependent on the quality of reference data source or annotated corpus.

Table 2 summarizes the works that deal with solutions for the geoparsing step.

4.3. Reference resolution solutions

We classify the methods for the reference resolution in three categories, following Buscaldi and Rosso (2008): *map-based* approaches; *knowledge-based* approaches; and *data driven* or *supervised* methods. We propose expanding the scope of map-based approaches, which involve geometries and distances, to include other GIS data, techniques and functions. If more complex geometric representations are available, methods can move beyond using simple positions and distances to use topological relationships for disambiguation (Fig. 3).

Methods and techniques classified as map-based approaches are those that use some geometric algorithms or topological functions, such as *disjoint*, *union*, *interception*, and others. Knowledge-based approaches are based on the hypothesis that toponyms appearing together in text are related to each other, and that this relation can be extracted from gazetteers and knowledge bases such as Wikipedia (Habib et al., 2013). Supervised methods are those based on standard machine learning techniques.

Fig. 3 shows the proposed classification for reference resolution approaches. Reference resolution proposals are very similar to each other. They usually distinguish themselves only on the geometric algorithm or topological function used.

The rationale behind using proximity as the basis of heuristics for disambiguation in *map-based* approaches is the well-known First Law of Geography (Tobler, 1970), i.e., the assumption that "everything is related to everything else, but near things are more related than distant things". The degree of belief in the semantics (e.g."Calgary, Alberta, Canada" or "Calgary, Zimbabwe") of an ambiguous toponym (e.g. "Calgary") can be measured based on vicinity measurements. In other words, the toponym is resolved to the place that is closest to other resolved toponyms (Shi and Barker, 2011).

For instance, Smith and Crane (2001) use the concepts of centroid in a map and calculate the distance of the centroid to the

[10] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html
[11] http://nlp.stanford.edu
[12] https://developers.google.com/maps/documentation/geocoding

Table 2
Solutions to the Geoparsing step.

	Lookup-based	Rule-based	Supervised
Language dependent	Alexopoulos et al. (2013), Amitay et al. (2004), Borges et al. (2007), Borges et al. (2011), Chasin et al. (2013), Chen et al. (2010), Clough (2005), Pouliquen et al. (2006), Purves et al. (2007), Shi and Barker (2011), Zubizarreta et al. (2008)	Lieberman and Samet (2011), Silva et al. (2006), Twaroch et al. (2008), Woodruff and Plaunt (1994)	
Language independent		Pouliquen et al. (2004)	Habib et al. (2013), Nissim et al. (2004)

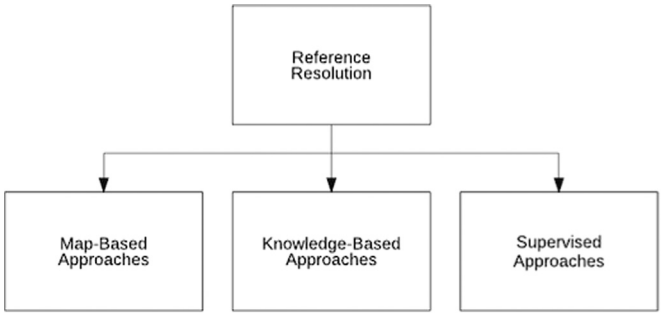

Fig. 3. Reference resolution approaches.

candidates to disambiguate the toponyms. Vargas et al. (2012a, 2012b) use a similar function based on a polygon containing the unambiguous entities, and Zong et al. (2005) use a gazetteer as support.

Leidner et al. (2003) define a *disambiguating context* as a region in whose confines most unresolved toponyms become unique. This can be implemented using a GIS function. He also uses the "one reference per discourse" heuristic.

A match against a gazetteer represents the simplest example of the *knowledge-based* approach (Chen et al., 2010; Pouliquen et al., 2004). Other simple examples try to use some semantic information present on the external knowledge source to disambiguate the references, such as population information (Rauch et al., 2003) or information about the administrative region and country names related to the ambiguous toponyms (Olligschlaeger and Hauptmann, 1999). Most work on reference resolution relies on using heuristics and custom-built rules which are knowledge source approaches since they use some gazetteer or database as support (DeLozier et al., 2015). For instance, Amitay et al. (2004) use a sequential algorithm that applies several heuristics to each ambiguous toponym. The first one looks for tokens in the vicinity and checks if they are able to uniquely qualify the toponyms; the second one solves the ambiguity by assuming as correct the geographic entity with the largest population. Next, the "one reference per discourse" heuristics is used, followed by another heuristic based on the notion of a disambiguating context. As opposed to Leidner et al. (2003), Amitay et al. (2004) use a administrative boundary hierarchy form external sources as a disambiguating context, with no actual map data involved.

Some works try to assign values, scores or probabilities to disambiguation candidates. This idea can be combined with some heuristics and/or classification rules (Alexopoulos et al., 2013; Li et al., 2006; Silva et al., 2006; Volz et al., 2007), or make use of variations of well-know algorithms, such as the GeoRank (Zhang et al., 2012) which uses the same voting process as in PageRank. Clough (2005) and Purves et al. (2007) explore the hierarchy of the external source to provide a default sense to an ambiguous toponym.

Other proposals use an external knowledge source and also include some algorithms that work on the string close to the ambiguous toponym in the text. Buscaldi and Rosso (2008) use WordNet[13] and a conceptual-density-based approach, selecting a context for each ambiguous toponym and building a sub-hierarchy that maximizes the conceptual density that helps to choose the correct term for the toponym. Li et al. (2002, 2003) also use this idea of using co-occurring words and build a data structure, in this case a graph, that allows to calculate the best match for the ambiguous toponym. Other works also deal with the core idea of using an external knowledge source and local context (Sobhana, 2012; Wang et al., 2010).

An unusual approach, by Andogah et al. (2012), proposes first to define the geographic scope of the document and then to use this scope with other heuristics to disambiguate the toponyms, only for geo-geo ambiguity. To solve the geo/non-geo cases, they use the Alias-i LingPipe[14] NER tool.

In the *supervised* approach, most of the works follow the usual Machine Learning approach: build a training set composed by disambiguated toponyms (in a manual or automatic way) and then run a standard learning algorithm, such as Naïve Bayes classifiers (Smith and Mann, 2003), Bayesian classifiers (Adelfio and Samet, 2013), Random Forests (Lieberman and Samet, 2012) or clustering methods (Habib and van Keulen, 2012). Instead of using unambiguous toponym to train the machine learning algorithms, some works use indirect supervision to create training data from links and Wikipedia annotations (Santos et al., 2014; Speriosu and Baldridge, 2013)

A gazetteer-based statistical classifier is used by Garbin and Mani (2005). The gazetteer is used to match the toponyms found in the text and also to train the machine learner over some features, including the class of all the toponyms in the document, i.e., national capital, political region or populated place, and the terms within three words from the ambiguous toponym. The system determines the class of the ambiguous toponyms and uses some preferences to disambiguate them.

A comparison between a map-based approach and a knowledge-based approach was made by Buscaldi and Rosso (2008). Both methods use the georeferenced version of WordNet. The first one calculates centroids considering each ambiguous candidate and unambiguous places. The centroid with the smallest accumulated distance is used to identify a preferred candidate. The second one exploits the structure of WordNet to solve ambiguity. They conclude that the knowledge-based approach was better when a small context was used, such as a sentence or a paragraph, and the map-based approach obtained best results when the context was the whole document.

The number of works that deal with the reference resolution problem is very large. However, as in the case of geoparsing, comparing all these methods is difficult, even though there are many similarities. Most solutions require an external knowledge source. As a result, the quality of the external dataset is crucial to achieve better results in reference resolution. For instance, according to Vasardani et al. (2013), the nature of each gazetteers'

[13] http://wordnet.princeton.edu
[14] http://alias-i.com/lingpipe

core components may create problems such as the lack of unofficial and vernacular names and the absence of a common structure among the gazetteers.

There are some efforts in creating and enriching these external sources. Machado et al. (2011) proposed an ontological gazetteer that includes geographic elements such as spatial relationships, concepts and terms related to places. Moura and Davis (2014) proposed the creation of a gazetteer from the integration of linked data sources, and found several issues with the quality of the data in such reference databases, including incompleteness and classification errors. Problems that derive from the quality of reference data have been reported for address geocoding problems (Hart and Zandbergen, 2013; Zandbergen, 2011; Ahlers, 2013), which are quite similar to GSR as to their methodology. Such works confirm that the accuracy of the results is fundamentally dependent on the completeness and accuracy of the reference dataset, although address geocoding uses a curated set of valid addresses or address ranges per thoroughfare segment (Davis and Fonseca, 2007). Nevertheless, the influence of the quality, completeness, level of detail and nature of the reference data source for these problems has not been assessed so far.

Table 3 summarizes the works that deal with the reference resolution step.

4.4. Solutions for grounding references

Solutions for grounding references are classified according to the form in which the geographic scope is generated and presented to the users and systems instead of considering the algorithms and techniques used to find the solution. References can be grounded as a one or more pairs of geographic coordinates. Besides that, another aspect that must be considered is the granularity of the solution, e.g., if the solution will be presented at the country, state or city level. Fig. 4 presents a taxonomy to organize these types of grounding.

The results of the grounding references step can be divided into those that consider a single place as the geographic scope of the document, and those that can generate multiple places to compose the scope. The latter can be divided into results that use some structure to inform the scope (structure) and the results that not reflect any kind of structure (unstructured).

The single place category includes techniques that consider the most representative place to be the geographic scope. It also uses some kind of generalization, grouping the places found in the document in a single place higher in a geographic or administrative hierarchy. This category includes the simplest alternatives, which just consider a single pair of coordinates to represent the geographic scope of documents (Amitay et al., 2004; Borges et al., 2011; Buyukkokten et al., 1999; Ding et al., 2000; McCurley, 2001; Woodruff and Plaunt, 1994). Other alternatives (Silva et al., 2006; Wang et al., 2005) use a data structure to calculate the importance of the toponyms identified in the geoparsing step, but they consider only the most important toponym found as the geographic scope.

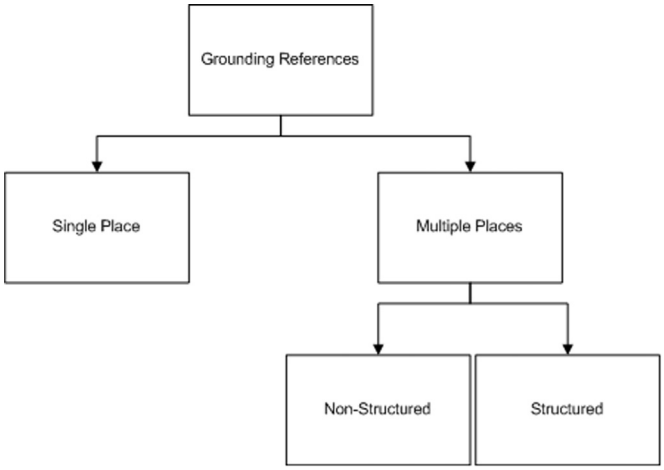

Fig. 4. Grounding references.

Table 4
Forms of the Geographic Scope.

Single place	Multiple places	
	Non-Structured	Structured
Amitay et al. (2004), Borges et al. (2011), Buyukkokten et al. (1999), Ding et al. (2000), McCurley (2001), Silva et al. (2006), Wang et al. (2005), Woodruff and Plaunt (1994)	Alexopoulos et al. (2013), Vargas et al. (2012a), Vargas et al. (2012b), Zong et al. (2005)	Andogah et al. (2012), Campelo and Baptista (2008), Chen et al. (2010), Zhang et al. (2012), Zubizarreta et al. (2008)

Most proposals that generate multiple places as a result define some data structure to organize them, such as a tree (Chen et al., 2010; Zhang et al., 2012; Campelo and Baptista, 2008) or a graph (Zubizarreta et al., 2008). Trees are used to represent a hierarchy of spatial subdivisions, so that parent nodes spatially contain their children. Moving up in the tree, it is possible to go from individual references to a single all-encompassing one that serves as a general result. Andogah et al. (2012) propose a particular data structure based on the assumption that "places of the same type or under the same administrative jurisdiction or adjacent to each other are more likely to be mentioned in a given discourse unit". Another alternative is a simple list of locations related to the text (Alexopoulos et al., 2013; Vargas et al., 2012b; Zong et al., 2005), in no particular ordering, which we group as non-structured approaches. Table 4 summarizes the forms in which the geographic scope is generated and presented.

Table 3
Solutions to the reference resolution step.

Map-based	Knowledge-Based	Supervised
Leidner et al. (2003), Shi and Barker (2011), Smith and Crane (2001), Vargas et al. (2012a), Vargas et al. (2012b), Zong et al. (2005)	Alexopoulos et al. (2013), Amitay et al. (2004), Andogah et al. (2012), Buscaldi and Rosso (2008), Chen et al. (2010), Clough (2005), DeLozier et al. (2015), Li et al. (2002), Li et al. (2003), Li et al. (2006), Olligschlaeger and Hauptmann (1999), Pouliquen et al. (2004), Purves et al. (2007), Rauch et al. (2003), Silva et al. (2006), Sobhana (2012), Volz et al. (2007), Wang et al. (2010), Zhang et al. (2012)	Adelfio and Samet (2013), Garbin and Mani (2005), Habib and van Keulen (2012), Lieberman and Samet (2012), Santos et al. (2014), Smith and Mann (2003), Speriosu and Baldridge (2013)

5. Conclusions and future work

The goal of this survey was to provide a comprehensive and structured view on contributions to the geographic scope resolution problem. We presented a definition of the GSR problem, dividing the problem in three steps: *geoparsing*, *reference resolution* and *grounding references*. Solutions to the first two steps were organized according to the method used, and proposals to the third step were organized according to the type of output produced. We would like to emphasize that the classes in each group of solutions are not exclusive, since in more complex solutions one of the steps might include characteristics of more than one class. In this way, some works could be classified in more than one class. Even though the summary tables classify each work according to its main focus, text in each section indicates works that cover other aspects of the problem as well.

One of the characteristics that stood out analyzing the solutions found in the literature was the use of external reference sources. The most common source used was a gazetteer, both for the global solutions and for geoparsing and reference resolution. Even some methods that rely on machine learning algorithms and heuristics use a gazetteer. The quality and the coverage of the gazetteer have an important role in the quality of the solution.

We also presented a list of related application areas that can benefit from the determination of the geographic scope of documents. Such areas range from search engines to social media, including social network user location inference and document classification according to spatial criteria. It should be noted that the list of applications and areas is not exhaustive. There are other works that only slightly differ from the subject of this article, but are closely related. For example, Wang et al. (2005) deal with three different kinds of geographic scope (provider, content and serving). Gravano et al. (2003) focus on determining the geographic scope of queries considering the locality issue, i.e., if the query is global or local. Anastácio et al. (2009) also deal with the locality issue to perform document classification. Quercini et al. (2010) highlight the difference between the geographic scope of a document and the geographic scope of readers; the former is related to the main geographic area of interest of a document, while the latter regards the audience that is interested in news about a specific place, i.e., a place geographically close to the reader.

Although the problem has been studied for some years and many solutions have been proposed in the literature, there are still several inconsistencies, mostly regarding terminology variations of the concepts related to the GSR problem. We suggested the use of some terms and made references to alternative uses.

In short, the main contributions of this survey are a proposal for standardizing the definition of the problem and related concepts, along with the classification of the solutions found in the literature to each step of the problem according to characteristics observed in the method.

Numerous research challenges and opportunities still exist, requiring novel solutions to the GSR problem and advances in related areas. First, the scope of a document can be semantically broader than simply a set of locations referenced in the text. As a result, the scope embeds part of the semantics of the document resulting in a *geographic semantic scope* that includes not only the explicitly mentioned locations but also the meaning that these references have to the readers of the documents. Second, the creation, maintenance and curation of reference data sources, such as gazetteers, is an essential part of the work in GIR, and should be much more active as a research topic in the future. Finally, it is difficult to compare the existing solutions directly to one another, especially due to the lack of annotated corpora (Wallgrün et al., 2014). Most of the works reviewed in this paper created their own benchmarking data, targeted to their methodologies and solutions.

Acknowledgments

This work was partially supported by FAPEMIG (grant CEX-PPM-00679/15) and CNPq (grants 303532/2015-7, 459818/2014-2 and 401822/2013-3), Brazilian agencies in charge of fostering research and development.

References

Adelfio, M.D., Samet, H., 2013. Geowhiz: toponym resolution using common categories. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL'13, ACM, New York, NY, USA, pp. 532–535. http://dx.doi.org/10.1145/2525314.2525321.

Ahlers, D., 2013. Assessment of the accuracy of Geonames gazetteer data, In: Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR '13, ACM, New York, NY, USA, pp. 74–81. http://dx.doi.org/10.1145/2533888.2533938.

Alencar, R.O., Davis Jr., C.A., 2011. Advancing Geoinformation Science for a Changing World, vol. 1, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 461–477 (Ch. Geotagging Aided by Topic Detection with Wikipedia). http://dx.doi.org/10.1007/978-3-642-19789-5_23.

Alencar, R.O., Davis Jr., C.A., Gonçalves, M.A., 2010. Geographical classification of documents using evidence from Wikipedia, In: Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10, ACM, New York, NY, USA, pp. 12:1–12:8. http://dx.doi.org/10.1145/1722080.1722096.

Alexopoulos, P., Ruiz, C., 2012. Optimizing geographical entity and scope resolution in texts using non-geographical semantic information. In: Proceedings of the 6th International Conference on Advances in Semantic Processing, SEMAPRO 2012, Think Mind, Barcelona, Spain, pp. 65–70. ⟨https://www.thinkmind.org/download.php?articleid=semapro_2012_3_40_50081⟩(accessed: 2016-05-30).

Alexopoulos, P., Ruiz, C., Villazon-Terrazas, B., Gómez-Pérez, J.M., 2013. KLocator: an ontology-based framework for scenario-driven geographical scope resolution. Int. J. Adv. Intell. Syst. 6 (3–4), 177–187, accessed: 2016-05-30. ⟨http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.683.4748&rep=rep1&type=pdf⟩.

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, NY, USA, pp. 273–280. http://dx.doi.org/10.1145/1008992.1009040.

Anastácio, I., Martins, B., Calado, P., 2009. Progress in Artificial Intelligence: 14th Portuguese Conference on Artificial Intelligence, EPIA 2009, Aveiro, Portugal, October 12–15, 2009. In: Proceedings, Lecture Notes in Computer Science, vol. 5816, Springer, Berlin, Heidelberg, 2009, pp. 598–609 (Ch. classifying documents according to locational relevance). http://dx.doi.org/10.1007/978-3-642-04686-5.49.

Anastácio, I., Martins, B., Calado, P., 2009. A comparison of different approaches for assigning geographic scopes to documents, In: Proceedings of the 1st InForum-Simpósio de Informática, INForum '09, Lisbon, Portugal, pp. 285–296. ⟨http://xldb.fc.ul.pt/xldb/publications/Anastacio.etal:AComparisonOf:2009_document.pdf⟩(accessed: 2016-05-31).

Andogah, G., Bouma, G., Nerbonne, J., 2012. Every document has a geographical scope. Data Knowl. Eng 81–82, 1–20. http://dx.doi.org/10.1016/j.datak.2012.07.002.

Backstrom, L., Sun, E., Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, pp. 61–70. http://dx.doi.org/10.1145/1772690.1772698.

Baeza-Yates, R.A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Borges, K.A.V., Laender, A.H.F., Medeiros, C.B., Davis Jr., C.A., 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, NY, USA, pp. 31–36. http://dx.doi.org/10.1145/1316948.1316957.

Borges, K.A.V., Davis Jr., C.A., Laender, A.H.F., Medeiros, C.B., 2011. Ontology-driven discovery of geospatial evidence in web pages. GeoInformatica 15 (4), 609–631. http://dx.doi.org/10.1007/s10707-010-0118-z.

Buscaldi, D., Rosso, P., 2008. A conceptual density-based approach for the disambiguation of toponyms. Int. J. Geogr. Inf. Sci. 22 (3), 301–313. http://dx.doi.org/10.1080/13658810701626251.

Buscaldi, D., Rosso, P., 2008. Map-based vs. knowledge-based toponym disambiguation. In: Proceedings of the 2nd International Workshop on Geographic Information Retrieval, GIR '08, ACM, New York, NY, USA, pp. 19–22. http://dx.doi.org/10.1145/1460007.1460011.

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., Shivakumar, N., 1999. Exploiting geographical location information of web pages. In: ACM SIGMOD Workshop on The Web and Databases (WebDB'99), pp. 91–96. ⟨http://ilpubs.stanford.edu:8090/415/1/1999-5.pdf⟩(accessed: 2016-05-30).

Campelo, C.E.C., Baptista, C.S., 2008. Geographic scope modeling for Web documents. In: Proceedings of the 2nd International Workshop on Geographic Information Retrieval, GIR '08, ACM, New York, NY, USA, pp. 11–18. http://dx.doi.org/10.1145/1460007.1460010.

Cardoso, N., Silva, M.J., Santos, D., 2008. Handling implicit geographic evidence for geographic IR. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, ACM, New York, NY, USA, 2008, pp. 1383–1384. http://dx.doi.org/10.1145/1458082.1458291.

Cardoso, N., 2011. Evaluating geographic information retrieval. SIGSPATIAL Spec. 3 (2), 46–53. http://dx.doi.org/10.1145/2047296.2047307.

Chasin, R., Woodward, D., Witmer, J., Kalita, J., 2013. Extracting and displaying temporal and geospatial entities from articles on historical events. Comput. J. 57 (3), 403–426. http://dx.doi.org/10.1093/comjnl/bxt112.

Chen, M., Lin, X., Zhang, Y., Wang, X., Yu, H., 2010. Assigning geographical focus to documents. In: 18th Conference on Geoinformatics, pp. 1–6. http://dx.doi.org/10.1109/GEOINFORMATICS.2010.5567598.

Clough, P., Sanderson, M., Joho, H., 2004. Extraction of Semantic Annotations from Textual Web Pages, Technical Report D15 6201, University of Sheffield, SPIRIT Project (EU IST-2001-35047). ⟨http://www.geo-spirit.org/publications/SPIRIT_WP6_D15_geomarkup_revised_FINAL.pdf⟩(accessed: 2016-05-30).

Clough, P., 2005. Extracting metadata for spatially-aware information retrieval on the internet. In: Proceedings of the 2005 Workshop on Geographic Information Retrieval, GIR '05, ACM, New York, NY, USA, pp. 25–30. http://dx.doi.org/10.1145/1096985.1096992.

Curran, J.R., Clark, S., 2003. Language independent NER using a maximum entropy tagger. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 – Volume 4, CONLL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 164–167. http://dx.doi.org/10.3115/1119176.1119200.

Davis Jr., C.A., Fonseca, F.T., 2007. Assessing the certainty of locations produced by an address geocoding system. GeoInformatica 11 (1), 103–129. http://dx.doi.org/10.1007/s10707-006-0015-7.

Davis Jr., C.A., Pappa, G.L., Oliveira, D.R.R., Arcanjo, F.L., 2011. Inferring the location of Twitter messages based on user relationships. Trans. GIS 15 (6), 735–751. http://dx.doi.org/10.1111/j.1467-9671.2011.01297.x.

Delboni, T.M., Borges, K.A.V., Laender, A.H.F., Davis Jr., C.A., 2007. Semantic expansion of geographic Web queries based on natural language positioning expressions. Trans. GIS 11 (3), 377–397. http://dx.doi.org/10.1111/j.1467-9671.2007.01051.x.

DeLozier, G., Baldridge, J., London, L., 2015. Gazetteer-independent toponym resolution using geographic word profiles. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, AAAI'15, AAAI Press, Austin, Texas, USA, pp. 2382–2388. ⟨http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9823/9530⟩(accessed: 2016-05-30).

Ding, J. Gravano, L., Shivakumar, N., 2000. Computing geographical scopes of Web resources. In: Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 545–556 (accessed: 2016-05-30). ⟨http://www.cs.columbia.edu/gravano/Papers/2000/vldb00.pdf⟩.

Drymonas, E., Pfoser, D., 2010. Geospatial route extraction from texts. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics, DMG '10, ACM, New York, NY, USA, pp. 29–37. http://dx.doi.org/10.1145/1869890.1869894.

Fu, G., Jones, C.B., Abdelmoty, A.I., 2005. Building a geographical ontology for intelligent spatial search on the Web. In: Hamza, M.H. (Ed.), Proceedings of IASTED International Conference on Databases and Applications, DBA 2005, IASTED/ACTA Press, Anaheim, CA, USA, pp. 167–172. ⟨http://www.geo-spirit.com/publications/geoontology_DBA.pdf⟩(accessed: 2016-05-31).

Fujiwara, Y., Sakurai, Y., Kitsuregawa, M., 2009. Fast likelihood search for hidden markov models. ACM Trans. Knowl. Discov. Data (TKDD) 3 (4). http://dx.doi.org/10.1145/1631162.1631166, 18:1–18:37.

Garbin, E., Mani, I., 2005. Disambiguating toponyms in news. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370. http://dx.doi.org/10.3115/1220575.1220621.

Goldberg, D.W., Wilson, J.P., Knoblock, C.A., 2007. From text to geographic coordinates: the current state of geocoding. URISA J. (J. Urban Reg. Inf. Assoc.) 19 (1), 33–47, accessed: 2016-05-30. ⟨https://www.researchgate.net/publication/281748610_From_text_to_geographic_coordinates_The_current_state_of_geocoding⟩.

Gouvêa, C., Loh, S., Garcia, L.F.F., Fonseca, E.B., Wendt, I., 2008. Discovering location indicators of toponyms from news to improve gazetteer-based geo-referencing. In: Carvalho, M.T. M., Casanova, M.A., Gattass, M., Vinhas, L. (Eds.), Proceedings of the X Brazilian Symposium on GeoInformatics, GeoInfo 2008, SBC, Porto Alegre, RS, Brazil, pp. 51–62. ⟨http://www.geoinfo.info/geoinfo2008/papers/p13.pdf⟩(accessed: 2016-05-31).

Gravano, L., Hatzivassiloglou, V., Lichtenstein, R., 2003. Categorizing Web queries according to geographical locality. In: Proceedings of the 12th International Conference on Information and Knowledge Management, CIKM '03, ACM, New York, NY, USA, pp. 325–333. http://dx.doi.org/10.1145/956863.956925.

Habib, M.B., van Keulen, M., 2011. Named entity extraction and disambiguation: The reinforcement effect. In: Proceedings of the 5th International Workshop on Management of Uncertain Data, MUD 2011, Seattle, USA, CTIT Workshop Proceedings Series, vol. WP11-02, Centre for Telematics and Information Technology, University of Twente, Enschede, pp. 9–16, Accessed: 2016-05-31. ⟨http://www.ctit.utwente.nl/library/proceedings/wp1102.pdf⟩.

Habib, M.B., van Keulen, M., 2012. Web Engineering. In: Proceedings of 12th International Conference, ICWE 2012, Berlin, Germany, July 23–27, 2012. Springer, Berlin, Heidelberg, pp. 439–443. http://dx.doi.org/10.1007/978-3-642-31753-8_39 (Ch. Improving toponym extraction and disambiguation using feedback loop).

Habib, M.B., van Keulen, M., 2013. A hybrid approach for robust multilingual toponym extraction and disambiguation. In: Kaopotek, M.A., Koronacki, J., Marciniak, M., Mykowiecka, A., Wierzchon, S.T. (Eds.), Language Processing and Intelligent Information Systems, Lecture Notes in Computer Science, Vol. 7912, Springer, Berlin, Heidelberg, pp. 1–15. http://dx.doi.org/10.1007/978-3-642-38634-3.1.

Hart, T.C., Zandbergen, P.A., 2013. Reference data and geocoding quality: examining completeness and positional accuracy of street geocoded crime incidents. Policing: Int. J. Police Strateg. Manag. 36 (2), 263–294. http://dx.doi.org/10.1108/13639511311329705.

Hearst, M.A., 1998. Support vector machines. IEEE Intell. Syst. 13 (4), 18–28. http://dx.doi.org/10.1109/5254.708428.

Hill, L.L., 2000. Research and Advanced Technology for Digital Libraries. In: Proceedings of 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 280–290. http://dx.doi.org/10.1007/3-540-45268-0_26 (Ch. Core elements of digital gazetteers: placenames, categories, and footprints).

Hill, L.L., 2006. Georeferencing: The Geographic Associations of Information. The MIT Press, ISBN: 978-0-262-08354-X.

Jones, C.B., Purves, R.S., 2008. Geographical information retrieval. Int. J. Geogr. Inf. Sci. 22 (3), 219–228. http://dx.doi.org/10.1080/13658810701626343.

Jones, C.B., Purves, R.S., 2009. Encyclopedia of Database Systems. Springer, US, Boston, MA, pp. 1227–1231. http://dx.doi.org/10.1007/978-0-387-39940-9_177, Ch. Geographic information retrieval.

Leidner, J.L., Lieberman, M.D., 2011. Detecting geographical references in the form of place names and associated spatial natural language. SIGSPATIAL Spec. 3 (2), 5–11. http://dx.doi.org/10.1145/2047296.2047298.

Leidner, J.L., Sinclair, G., Webber, B., 2003. Grounding spatial named entities for information extraction and question answering. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, HLT-NAACL-GEOREF '03, vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 31–38. http://dx.doi.org/10.3115/1119394.1119399.

Leidner, J.L., 2007. Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names (Ph.D. thesis), University of Edinburgh (June).

Leidner, J.L., 2008. Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal-Publishers, ISBN: 978-1-58112-384-5.

Leveling, J., Hartrumpf, S., Veiel, D., 2006. Using semantic networks for geographic information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G., Kluck, M., Magnini, B., de Rijke, M. (Eds.), Accessing Multilingual Information Repositories. Lecture Notes in Computer Science, vol. 4022, Springer, Berlin, Heidelberg, pp. 977–986. http://dx.doi.org/10.1007/11878773.109.

Li, H., Srihari, R.K., Niu, C., Li, W., 2002. Location normalization for information extraction. In: Proceedings of the 19th International Conference on Computational Linguistics, COLING '02, vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–7. http://dx.doi.org/10.3115/1072228.1072355.

Li, H., Srihari, R.K., Niu, C., Li, W., 2003. Infoxtract location normalization: a hybrid approach to geographic references in information extraction. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, HLT-NAACL-GEOREF '03, vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 39–44. http://dx.doi.org/10.3115/1119394.1119400.

Li, Y., Moffat, A., Stokes, N., Cavedon, L., 2006. Exploring probabilistic toponym resolution for geographical information retrieval. In: Purves, R., Jones, C. (Eds.), Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, SIGIR 2006, Department of Geography, University of Zurich, Seattle, WA, USA. ⟨http://www.geo.unizh.ch/rsp/gir06/papers/individual/li.pdf⟩(accessed: 2016-05-30).

Lieberman, M.D., Samet, H., 2011. Multifaceted toponym recognition for streaming news. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11. ACM, New York, NY, USA, pp. 843–852. http://dx.doi.org/10.1145/2009916.2010029.

Lieberman, M.D., Samet, H., 2012. Supporting rapid processing and interactive map-based exploration of streaming news. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12. ACM, New York, NY, USA, pp. 179–188. http://dx.doi.org/10.1145/2424321.2424345.

Lieberman, M.D., Samet, H., 2012. Adaptive context features for toponym resolution in streaming news. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12. ACM, New York, NY, USA, pp. 731–740. http://dx.doi.org/10.1145/2348283.2348381.

Lieberman, M.D., Samet, H., Sankaranarayanan, J., 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In: Proceedings of the IEEE 26th International Conference on Data Engineering (ICDE 2010), IEEE, Long Beach, California, USA, pp. 201–212. http://dx.doi.org/10.1109/ICDE.2010.5447903.

Lim, E.-P., Goh, D.H.-L., Liu, Z., Ng, W.-K., Khoo, C.S.-G., Higgins, S.E., 2002. G-portal: a map-based digital library for distributed geospatial and georeferenced resources. In: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '02. ACM, New York, NY, USA, pp. 351–358. http://dx.doi.org/10.1145/544220.544307.

Luo, J., Joshi, D., Yu, J., Gallagher, A., 2011. Geotagging in multimedia and computer vision – a survey. Multimed. Tools Appl. 51 (1), 187–211. http://dx.doi.org/10.1007/s11042-010-0623-y.

Machado, I.M.R., Alencar, R.O., Campos Jr., R.O., Davis Jr., C.A., 2011. An ontological gazetteer and it's application for place name disambiguation in text. J. Braz. Comput. Soc. 17 (4), 267–279. http://dx.doi.org/10.1007/s13173-011-0044-4.

Mark, D., 1989. Cognitive image-schemata for geographic information: Relations to user views and GIS interfaces. In: Proceedings of GIS/LIS 89': Annual Conference and Exposition, vol. 2, Orlando, Florida, USA, pp. 551–560.

Martins, B., Cardoso, N., Chaves, M.S., Andrade, L., Silva, M.J., 2007. The University of Lisbon at GeoCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (Eds.), Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science, vol. 4730, Springer, Berlin, Heidelberg, pp. 986–994. http://dx.doi.org/10.1007/978-3-540-74999-8.127.

McCurley K.S., 2001. Geospatial mapping and navigation of the Web. In: Proceedings of the 10th International Conference on World Wide Web, WWW '01. ACM, New York, NY, USA, pp. 221–229. http://dx.doi.org/10.1145/371920.372056.

Morimoto, Y., Aono, M., Houle, M.E., McCurley, K.S., 2003. Extracting spatial knowledge from the Web. In: Symposium on Applications and the Internet, pp. 326–333. http://dx.doi.org/10.1109/SAINT.2003.1183066.

Moura, T.H.V.M., Davis Jr., C.A., 2014. Integration of linked data sources for gazetteer expansion. In: Proceedings of the 8th ACM SIGSPATIAL Workshop on Geographic Information Retrieval, GIR '14. ACM, New York, NY, USA, pp. 5:1–5:8. http://dx.doi.org/10.1145/2675354.2675357.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvist. Investig. 30, 3–26. http://dx.doi.org/10.1075/li.30.1.03nad, Accessed: 2016-05-30. ⟨http://nlp.cs.nyu.edu/sekine/papers/li07.pdf⟩.

Nissim, M., Matheson, C., Reid, J., 2004. Recognising geographical entities in Scottish historical documents. In: Purves, R., Jones, C. (Eds.). Proceedings of the Workshop On Geographic Information Retrieval, SIGIR 2004, Sheffield, England, July 25, 2004, Department of Geography, University of Zurich, Sheffield, England. ⟨https://www.ltg.ed.ac.uk/np/publications/ltg/papers/Nissim2004Recognising.pdf⟩(accessed: 2016-05-30).

Olligschlaeger, A.M., Hauptmann, A.G., 1999. Multimodal information systems and GIS: the informedia digital video library. In: 1999 ESRI User Conference, Environmental Systems Research Institute (ESRI) Inc., San Diego, California, USA. ⟨http://www.informedia.cs.cmu.edu/documents/ESRI99.html⟩(accessed: 2016-05-30).

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The Pagerank Citation Ranking:

Bringing Order to the Web, Technical Report, Stanford University. ⟨http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf⟩(accessed: 2016-05-30).

Pouliquen, B., Steinberger, R., Ignat, C., Groeve, T., 2004. Geographical information recognition and visualization in texts written in various languages. In: Proceedings of the 2004 ACM Symposium on Applied Computing, SAC '04, ACM, New York, NY, USA, pp. 1051–1058. http://dx.doi.org/10.1145/967900.968115.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fuart, F., Zaghouani, W., Widiger, A., Forslund, A.-C., Best, C., 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy, pp. 53–58. ⟨https://arxiv.org/pdf/cs/0609065.pdf⟩(accessed: 2016-05-30).

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B., 2007. The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet. Int. J. Geogr. Inf. Sci. 21 (7), 717–745. http://dx.doi.org/10.1080/13658810601169840.

Quercini, G., Samet, H., Sankaranarayanan, J., Lieberman, M.D., Determining the spatial reader scopes of news sources using local lexicons. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10, ACM, New York, NY, USA, 2010, pp. 43–52. http://dx.doi.org/10.1145/1869790.1869800.

Rauch, E., Bukatin, M., Baker, K., 2003. A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, HLT-NAACL-GEOREF '03, vol. 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 50–54. http://dx.doi.org/10.3115/1119394.1119402.

Ribeiro Jr., S.S., Davis Jr., C.A., Oliveira, D.R.R., Meira Jr., W., Gonçalves, T.S., Pappa, G.L., 2012. Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '12, ACM, New York, NY, USA, pp. 5–11. http://dx.doi.org/10.1145/2442796.2442800.

Sanderson, M., Kohler, J., 2004. Analyzing geographic queries. In: Purves, R., Jones, C. (Eds.), Proceedings of the Workshop On Geographic Information Retrieval, (SIGIR 2004), Department of Geography, University of Zurich, Sheffield, England, 2004. ⟨http://marksanderson.org/publications/my_papers/GeoQueryAnalysis2004.pdf⟩(accessed: 2016-05-30).

Santos, J., Anastácio, I., Martins, B., 2014. Using machine learning methods for disambiguating place references in textual documents. GeoJournal 80 (3), 375–392. http://dx.doi.org/10.1007/s10708-014-9553-y.

Shi, G., Barker, K., 2011. Extraction of geospatial information on the Web for GIS applications. In: 10th IEEE International Conference on Cognitive Informatics Cognitive Computing, ICC*CC, IEEE, Banff, Alberta, Canada, pp. 41–48. http://dx.doi.org/10.1109/COGINF.2011.6016120.

Silva, M.J., Martins, B., Chaves, M.S., Afonso, A.P., Cardoso, N., 2006. Adding geographic scopes to web resources. Comput. Environ. Urban Syst. 30 (4), 378–399. http://dx.doi.org/10.1016/j.compenvurbsys.2005.08.003.

Smith, D.A., 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-Verlag, London, UK, UK, pp. 127–136.

Smith, D.A., Mann, G.S., 2003. Bootstrapping toponym classifiers. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, HLT-NAACL-GEOREF '03, vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 45–49. http://dx.doi.org/10.3115/1119394.1119401.

Sobhana, N., 2012. Enhancing retrieval of geological text using named entity disambiguation. Int. J. Emerg. Technol. Adv. Eng. 2 (1), 2250–2459, accessed: 2016-05-30. ⟨http://www.ijetae.com/files/Volume2Issue1/IJETAE_0112_48.pdf⟩.

Souza, L.A., Davis Jr., C.A., Borges, K.A.V., Delboni, T.M., Laender, A.H.F., The role of gazetteers in geographic knowledge discovery on the Web. In: Proceedings of the 3rd Latin American Web Congress, LA-WEB '05, IEEE Computer Society, Washington, DC, USA, 2005, p. 157. http://dx.doi.org/10.1109/LAWEB.2005.38.

Speriosu, M., Baldridge, J., 2013. Text-driven toponym resolution using indirect supervision. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, pp. 1466–1476. ⟨http://www.aclweb.org/anthology/P13-1144⟩ (accessed: 2016-05-30).

Teitler, B.E., Lieberman, M.D., 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, NY, USA, pp. 18:1–18:10. http://dx.doi.org/10.1145/1463434.1463458.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46 (2), 234–240. http://dx.doi.org/10.2307/143141.

Twaroch, F.A., Smart, P.D., Jones, C.B., 2008. Mining the Web to detect place names. In: Proceedings of the 2nd International Workshop on Geographic Information Retrieval, GIR '08, ACM, New York, NY, USA, pp. 43–44. http://dx.doi.org/10.1145/1460007.1460017.

Vaid, S., Jones, C.B., Joho, H., Sanderson, M., 2005. Spatio-textual indexing for geographical search on the Web. In: Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases, SSTD'05, Springer-Verlag, Berlin, Heidelberg, Germany, pp. 218–235. http://dx.doi.org/10.1007/11535331_13.

Vargas, R.N.P., Moura, M.F., Speranza, E.A., Rodriguez, E., Rezende, S.O., 2012. Discovering the spatial coverage of the documents through the Spatial CIM methodology. In: Proceedings of the 15th AGILE'2012 International Conference on Geographic Information Science, Avignon, France, April, pp. 181–186. ⟨http://sites.labic.icmc.usp.br/pub/solange/AGILE_Nathy_2012.pdf⟩(accessed: 2016-05-30).

Vargas, R.N.P., Moura, M.F., Speranza, E.A., Rodriguez, E., Rezende, S.O., 2012b. The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques. In: Geospatial Information and Documents. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 16. (Geo Doc 2012), PAKDD Workshop, Kuala Lumpur, Malaysia. ⟨http://www.lirmm.fr/geodoc2012/Proceedings/Proceedings_GeoDoc.pdf⟩(accessed: 2016-05-31).

Vasardani, M., Winter, S., Richter, K.-F., 2013. Locating place names from place descriptions. Int. J. Geogr. Inf. Sci. 27 (12), 2509–2532. http://dx.doi.org/10.1080/13658816.2013.785550.

Volz, R., Kleb, J., Mueller, W., 2007. Towards ontology-based disambiguation of geographical identifiers. In: Bouquet, P., Stoermer, H., Tummarello, G., Halpin, H., (Eds.). Proceedings of the WWW2007 Workshop i3: Identity, Identifiers, Identification, CEUR Workshop Proceedings, Banff, Canada. ⟨http://ceur-ws.org/Vol-249/submission_132.pdf⟩(accessed: 2016-05-30).

Wallgrün, J.O., Klippel, A., Baldwin, T., 2014. Building a corpus of spatial relational expressions extracted from Web documents. In: Proceedings of the 8th Workshop on Geographic Information Retrieval, GIR '14, ACM, New York, NY, USA, 2014, pp. 6:1–6:8. http://dx.doi.org/10.1145/2675354.2675702.

Wang, C., Xie, X., Wang, L., Lu, Y., Ma, W.-Y., 2005. Detecting geographic locations from Web resources. In: Proceedings of the 2005 Workshop on Geographic Information Retrieval, GIR '05, ACM, New York, NY, USA, pp. 17–24. http://dx.doi.org/10.1145/1096985.1096991.

Wang, X., Zhang, Y., Chen, M., Lin, X., Yu, H., Liu, Y., 2010. An evidence-based approach for toponym disambiguation. In: 2010 18th International Conference on Geoinformatics, pp. 1–7. http://dx.doi.org/10.1109/GEOINFORMATICS.2010.5567805.

Woodruff, A.G., Plaunt, C., 1994. GIPSY: automated geographic indexing of text documents. J. Am. Soc. Inf. Sci. – Spec. Issue: Spat. Inf. 45 (9), 645–655. http://dx.doi.org/10.1002/(SICI)1097-4571(199410)45:9⟨645::AID-ASI2⟩3.0.CO;2-8.

Yang, S., Kavanaugh, A., Kozievitch, N.P., Li, L.T., Srinivasan, V. Sheetz, S.D., Whalen, T., Shoemaker, D., Torres, R., Fox, E.A., 2011. CTRnet DL for disaster information services. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL '11, ACM, New York, NY, USA, pp. 437–438. http://dx.doi.org/10.1145/1998076.1998173.

Zandbergen, P.A., 2011. Influence of street reference data on geocoding quality. Geocarto Int. 26 (1), 35–47. http://dx.doi.org/10.1080/10106049.2010.537374.

Zhang, Q., Jin, P., Lin, S., Yue, L., 2012. Extracting focused locations for Web pages. In: Proceedings of the 2011 International Conference on Web-Age Information Management, WAIM'11, Springer-Verlag, Berlin, Heidelberg, pp. 76–89. http://dx.doi.org/10.1007/978-3-642-28635-37.

Zong, W., Wu, D., Sun, A., Lim, E.-P., Goh, D.H.-L., 2005. On assigning place names to geography related Web pages. In: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '05, ACM, New York, NY, USA, pp. 354–362. http://dx.doi.org/10.1145/1065385.1065464.

Zubizarreta, A., Fuente, P., Cantera, J.M., Arias, M., Cabrero, J., García, G., Llamas, C., Vegas, J., 2008. A georeferencing multistage method for locating geographic context in Web search. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, ACM, New York, NY, USA, pp. 1485–1486. http://dx.doi.org/10.1145/1458082.1458347.