An Ontological Gazetteer for Geographic Information Retrieval

Ivre Marjorie R. Machado[1], Rafael Odon de Alencar[1], Roberto de Oliveira Campos Junior[2,3], Clodoveu A. Davis Junior[1]

[1]Dep. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos 6627 – ICEx – 31270-010 – Belo Horizonte – MG – Brasil

[2]Prog. de Pós-graduação em Engenharia Elétrica – Pontifícia Universidade Católica de Minas Gerais (Puc Minas) – Belo Horizonte – MG – Brasil

[3]Serviço Federal de Processamento de Dados (SERPRO)

{ivre,odon,clodoveu}@dcc.ufmg.br,roberto.oliveira@ieee.org

Abstract. *The volume of spatial information on the Web grows daily, added to that, the problems to recognize references to spatial relationships and to deal with ambiguous names. This article presents a gazetteer, which has a structure different from conventional gazetteers. The ontological gazetteer will not only identify the names of places, but also record concepts and terms related to a place, as in an ontology in which concepts are the main places and features. A case study showed good results for detection of place names and inference implied by news Web sites based on content of ontological gazetteer.*

1. Introduction

The volume of information currently available on the Web is very large, and grows daily. Retrieving such information requires systems that are capable of understanding the needs of the users, locating 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 retrieval needs in many ways, but mostly in the form of sets of keywords submitted to a search engine. Previous work (Sanderson and Kohler 2004; Wang, Wang et al. 2005; Delboni, Borges et al. 2007; Backstrom, Kleinberg et al. 2008) 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 geographic context. For instance, if a set of places associated to a document can be determined, it would be possible to modify the document's 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, dictionary of place names (Hill 2000).

Current gazetteers are available on the Web, and 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 (a simple pair of coordinates indicating its location). With this kind of structure, gazetteers present several limitations, which make them harder to employ in information retrieval problems. Furthermore, gazetteer contents usually do not include intra-urban place names, such as street names, neighborhoods, landmarks or tourist attractions, and there are no resources with which to record and use the spatial relationship among its elements, other than estimating their proximity based on footprints. Gazetteers are notoriously hard to maintain and to expand, and as a result 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, 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 the limited structure, several Web-based geographic applications use information from gazetteers, as demonstrated by Goodchild and Hill (2008). 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*, 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 towards solving difficult problems such as place name disambiguation, geographic text classification, and geographic context recognition (Silva, Martins et al. 2004; Wang, Xie et al. 2005; Adriani and Paramita 2007; Overell and Ruger 2007).

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. Finally, Section 6 presents our conclusions and a rather extensive list of future work.

2. Related work

Hill (2000) 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 (i.e., the toponymical dictionaries usually found in atlases) gazetteers, and have been used as the basis for the development of the Alexandria Digital Library (ADL) Gazetteer. Since this pioneering initiative such basic structure has also been used in other Web-based gazetteer projects.

Uryupina (2003), Toral and Munoz (2006), and Popescu, Grefenstette et al. (2008) present proposals for populating and automatically maintaining gazetteers. These works extract data from the Wikipedia, which is a large knowledge base in different languages. Gouvêa, Loh et al (2008) propose a strategy for the identification of entities found in news texts, to be used in the development and updating of gazetteers. Alencar, Davis-Jr. et al. (2010) describe a strategy for classifying text into geographic categories through data extraction from Wikipedia to find evidence of place names in texts.

Lopez-Pellicer, Silva et al. (2010) present Geo-Net-PT 02, a geographic ontology of Portugal, an evolution of Geo-Net-PT 01 (Rodrigues, Chaves et al. 2006). 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.

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, Har'El et al. (2004) present Web-a-Where, a system that identifies geotags for Web pages with the support of a gazetteer. Souza, Davis-Jr et (2005) and Souza, Delboni et al. (2004) developed Locus, a geographic locator built around a gazetteer and based on a previously created ontology, OnLocus (Borges 2006; Borges, Laender et al. 2007). Overell and Ruger (2007) describe a model based on co-occurrence to solve the place name ambiguity problem, which uses a combination of heuristics and gazetteers.

Our work proposes changes in the structure of the gazetteer and demonstrate that it can be used in problems of Geographic Information Retrieval. We have used several sources to populate the gazetteer, including data from Wikipedia. This structure is an ontological construct, which enables the understanding and expansion semantics between entities. The *ontological gazetteer* is presented in the next section.

3. Ontological Gazetteer

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 the atlas each place name is associated to a generic type, a map number and a grid coordinate, in digital gazetteers a pair of geographic coordinates (lat-long) is used as a footprint. There are also known variants of each place name, such as abbreviations and popular names, as well as language-specific versions.

The place type comes from a previously determined hierarchy, which varies among gazetteers. For instance, the Alexandria Digital Library Gazetteer (ADL)[1] (Hill 2000) 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[2] gazetteer defines feature codes, with the first level consisting of nine classes, with a single level of further specialization. Other digital gazetteers include TGN[3] (Getty Thesaurus of Geographic Names) and GKB[4] (Global Knowledge Base).

[1] http://www.alexandria.ucsb.edu/

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

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

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

Previous works (Fu, Jones et al. 2005; Souza, Davis-Jr et al. 2005; Borges 2006; Borges, Laender et al. 2007) 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 know 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 (Gouvêa, Loh et al. 2008).

Souza, Davis-Jr et al. (2005) 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* (Borges 2006; Borges, Laender et al. 2007). 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 (Borges, Laender et al. 2007). In turn, the good performance of such indirect references in geographic information retrieval tasks suggested that a gazetteer might be much more helpful if it could record the various types of relationships that exist between places, going beyond the topology of geographic objects and allowing the inclusion of other types of semantic relationships. In order to implement this kind of semantically richer relationships, the gazetteer's design needs to include the flexible structure often found in ontology creation tools, such as Protégé[5], thus becoming what we call an *ontological gazetteer*, or *ontogazetteer*.

In order to accomplish this enlarged role, the ontogazetteer must be able to record various types of relationships 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. The ontogazetteer also must be able to record alternative names to a place as synonyms, also adapting the notions of hyponymy and hyperonymy for territorial subdivisions.

Another proposed enhancement for the ontogazetteer is the association of natural-language terms and expressions to each place. The idea is to improve the available information resources for performing typical geographic information retrieval tasks, such as disambiguation and geographic context recognition. An experimental procedure for obtaining these terms has been presented in (Alencar, Davis-Jr. et al. 2010), along with a classification procedure.

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

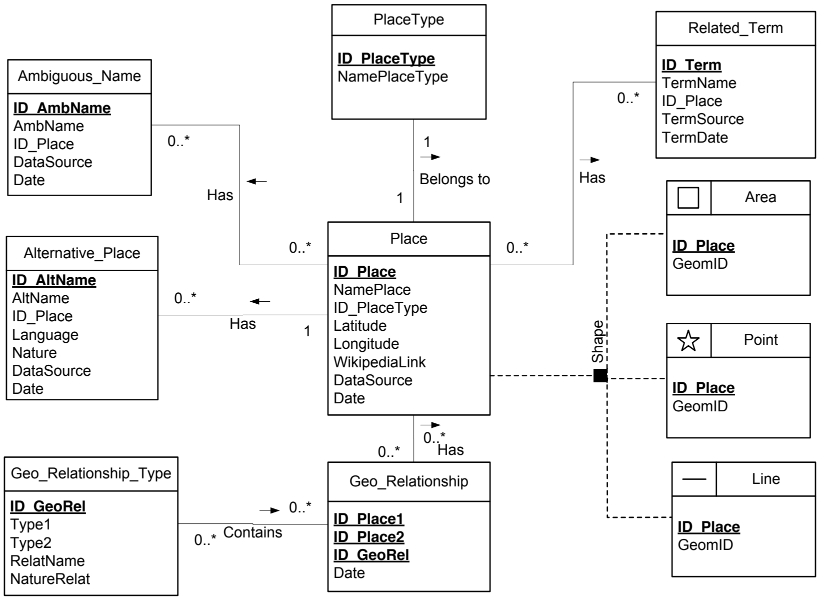


Figure 1. Gazetteer conceptual schema

Figure 1 presents the OMT-G (Borges, Davis-Jr et al. 2001) schema proposed for the ontological gazetteer. The schema represents all place names using the Place class. The Alternative_Place class maintains alternative names, abbreviations, acronyms, popular names and other variations. Each place belongs to a PlaceType; we initially based our place type definitions on the feature code thesaurus from ADL. Alternative names, abbreviations, acronyms and others are kept in the AlternativePlace class. Relationships between places are maintained by the GeoRelationship class; notice that two places can have various 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. In order to support disambiguation, there is a class that keeps known ambiguous place names (Ambiguous_Name). 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 (Alencar, Davis-Jr. et al. 2010) (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 “acarajé” 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 (Davis-Jr and Laender 1999). Places represented by more complex geometries will also have a point representation, as in current gazetteers.

The class diagram in Figure 1 was detailed and mapped to a geographic database. It is currently being populated, using existing geographic data and other gazetteers as primary sources. From the OnLocus ontology, we derived several types of geographic relationships between places. Special procedures and triggers have been created for each 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 (Jones and Purves 2008) 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 (Jones and Purves 2008). *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 (Amitay, Har'El et al. 2004; Teitler, Lieberman et al. 2008). Both geoparsing and geotagging require the recognition of geographic references found in text, and if this task is fulfilled adequately, the geographic context of the document can be established.

The ontogazetteer maintains lists of official and alternative 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 spatial queries during a GIR-related process. Therefore, applications can decide on the types of relationships that are to be considered, and which entities are to be taken into consideration. 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, the name "São Paulo" exists in 6,522 different GeoNames records. According to Smith and Crane (2001), 92% of TGN's toponyms are ambiguous. Several different types of ambiguities have been described in previous research (Amitay, Har'El et al. 2004; Volz, Kleb et al. 2007).

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 (Jones and Purves 2008). 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. 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 (Jones and Purves 2008). 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 (Delboni, Borges et al. 2007).

4.4. Spatial and textual indexing

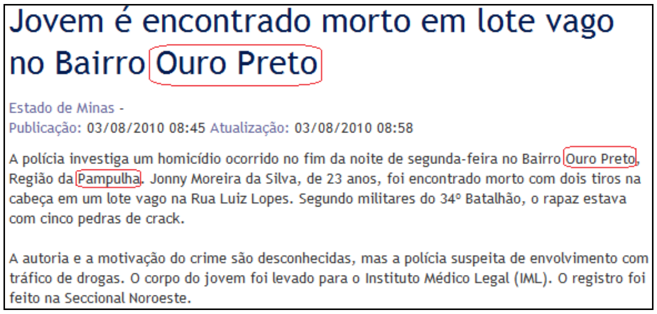
One of the techniques for indexing the contents of a text document is the creation of an inverted index file for the words contained in the document. This index provides, therefore, an association of each word to the list of documents that contain it. In the case of geographic references, this idea can 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 (Jones and Purves 2008). 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 (Jones and Purves 2008). Since the ontogazeteer 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. From footprints and geographic representations, proximity relationships can be determined, so that aspect can influence the ranking as well.

5. Case study

We exemplify the use of the proposed ontogazetteer with a case study. 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, detect the occurrence of place names, and infer the geographic context of each of them. We are able to recognize both explicitly and implicitly mentioned places. The latter are those whose relationship to the facts in the text is implied by the contents, but which are not directly mentioned.



Jovem é encontrado morto em lote vago no Bairro Ouro Preto

Estado de Minas -
Publicação: 03/08/2010 08:45 Atualização: 03/08/2010 08:58

A polícia investiga um homicídio ocorrido no fim da noite de segunda-feira no Bairro Ouro Preto, Região da Pampulha. Jonny Moreira da Silva, de 23 anos, foi encontrado morto com dois tiros na cabeça em um lote vago na Rua Luiz Lopes. Segundo militares do 34° Batalhão, o rapaz estava com cinco pedras de crack.

A autoria e a motivação do crime são desconhecidas, mas a polícia suspeita de envolvimento com tráfico de drogas. O corpo do jovem foi levado para o Instituto Médico Legal (IML). O registro foi feito na Seccional Noroeste.

Figure 3. Example (source: Uai – August 3, 2010)

Figure 3 presents a sample news text (in Portuguese) from the *Uai Minas*[6] news source, published August 3, 2010. Two place names have been identified, "Ouro Preto" and "Pampulha". In this case, "Ouro Preto" is ambiguous: the text refers to a neighborhood in Belo Horizonte, not to the famous historical city; nevertheless, both are obtained as candidate places. Consulting gazetteer data, other places related to the ones explicitly mentioned can be obtained, including "Belo Horizonte" (city), "Belo Horizonte Metropolitan" (micro-region), "Ouro Preto" (micro-region), "Metalurgical and Campos das Vertentes" (macro-region), and "Minas Gerais" (state). The latter places are implicit in the text. The text also includes a reference to a street ("Rua Luiz Lopes"), but the gazetteer currently does not include street data.

[6] http://www.uai.com.br/htmls/app/noticia173/2010/08/03/noticia_minas,i=172053/index.shtml

5.1. Creation of the news texts collection

The collection of news texts was performed from June to August 2010 (Table 1). 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 were collected, because most of the gazetteer data put together so far refer to this state. In order to ensure that, news were obtained from local- or state-related sections of the news sites.

Table 1. Web news sources

News Web	News Web Site	Local News Section Name	# Docs
Globominas	http://globominas.globo.com/	General News	139
O Tempo	http://www.otempo.com.br/	Latest News (cities)	75
Uai	http://www.uai.com.br/	Minas	71
Terra	http://www.terra.com.br	Latest News (Brasil)	11
		Total	**296**

5.2. Detection and inference of place names

After the news documents were collected, a pre-processing step removed stopwords (except for "de", "da(s)", "do(s)", which are quite common in Brazilian place names). 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 most of the explicitly mentioned names. Typically, names of places that are higher in a territorial hierarchy were found.

5.3. Experimental evaluation

In order to verify whether the place names recognized from the news documents (both explicitly and implicitly) were valid, a manual verification of each document was performed by a group of volunteers, composed by people with various backgrounds. A Web interface was developed[7], showing the text's title and body, along with a list of the implicit and explicit places, generated as described in the previous section. For each of these places, we asked the volunteers to determine the degree of its relationship to the news, using a scale with three levels: (0) unrelated, (1) slightly related, and (2) strongly related. Furthermore, the volunteer had the possibility of skipping the evaluation, if she did not know the place or if the determination could not be made from the existing information. We also included a text field so that volunteers could record observations or indicate difficulties and special cases. From the 296 news documents originally collected, we were able to find place names in 267, which were then used for this experiment. An average of 8,46 places were found per document.

[7] http://94.229.77.252/AvaliacaoCGN/

5.4. Experiment results

From the 267 news documents containing place names, 100% were verified by volunteers. Overall, 2,244 relationships between places and documents were evaluated, of which 72% were considered to be valid, and most of those were considered strong relationships (Table 2).

Table 2. Evaluation according to degree of relationship

Degree of relationship	Count	%
0- Unrelated	609	27,14
1- Slightly related	106	4,72
2- Strongly related	1.455	64,84
3- Indifferent	74	3,30
Total	**2.244**	**100,00**

Table 3. Evaluation according to type of place

Type	Total	Implicit refs	Avg.	Std.dev.
Macrorregion	184	171	0,36	0,75
Microrregion	298	129	0,98	0,94
River	2	0	1,00	1,41
Mesorregion	246	0	1,17	0,95
Municipality	446	35	1,50	0,85
State	341	233	1,61	0,75
Neighborhood	546	0	1,73	0,67
Highway	107	0	1,91	0,42

Table 3 shows the evaluation according to the type of place. The average and standard deviation columns refer to the degree of relationship, i.e., averages approaching 2 indicate consensus that there is a strong relationship between places of a type and the documents. We observed that places whose names are more widely known ranked better in the evaluation. Place types that belong to an administrative territorial hierarchy, such as macrorregion, mesorregion and microrregion were included mostly as implicit context, but people had difficulties recognizing their names. As an example, few volunteers apparently knew that the macrorregion in Minas Gerais to which Belo Horizonte belongs is called "Metalúrgica e Campos das Vertentes". Naturally, such names are hardly ever mentioned in journalistic texts.

We also noticed a high number of explicit mentions to neighborhoods, reinforcing the idea that gazetteers should cover intra-urban detail. Highways were also frequently cited, and from this fact we surmise that the relationships between places along a highway can be semantically very important. Notice also the high incidence of implicit references to the state, something that was expected, due to the fact that the sources of news were state-related sections of news sites. On the down side, the very small number of references to rivers indicates that the gazetteer's coverage of hydrography features is currently deficient.

6. Conclusions and Future Work

This paper proposed a new structure for gazetteers that seeks to diminish their limitations as components of geographic 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 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 (Alencar, Davis-Jr. et al. 2010).

A case study implemented a GIR task, namely the identification of geographic context 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. Placename recognition achieved good results, mainly for more the types of places that are more usually found in news, such as municipality, neighborhood, highway and state.

Future work includes developing an extension of the case study with a broader base of documents. 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.

7. Acknowledgements

This work was partially supported by the Brazilian research funding agencies CNPq and FAPEMIG.

References

Adriani, M. and M. L. Paramita (2007). Identifying location in indonesian documents for geographic information retrieval. In Proc. of the 4th GIR'07. Lisbon, Portugal, ACM.

Alencar, R. O., C. A. Davis-Jr., et al. (2010). Geographical classification of documents using evidence from Wikipedia. In Proc. of the 6th Workshop on GIR. Zurich, Switzerland, ACM.

Amitay, E., N. Har'El, et al. (2004). Web-a-where: geotagging web content. In Proc. of the 27th SIGIR Conf. on Research and Development in Information Retrieval. Sheffield, UK, ACM.

Backstrom, L., J. Kleinberg, et al. (2008). Spatial variation in search engine queries. Proceeding of the 17th International Conference on WWW. Beijing, China, ACM.

Borges, K. A. V. (2006). Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Belo Horizonte, UFMG. Doutorado: 195.

Borges, K. A. V., C. A. Davis-Jr, et al. (2001). "OMT-G: An object-oriented data model for geographic applications." GeoInformatica 5(3): 221-260.

Borges, K. A. V., A. H. F. Laender, et al. (2007). Discovering geographic locations in web pages using urban addresses. In Proc. of the 4th Workshop on GIR. Lisbon, Portugal, ACM.

Davis-Jr, C. A. and A. H. F. Laender (1999). Multiple representations in GIS: materialization through map generalization, geometric, and spatial analysis operations. In Proc. of the 7th ACM GIS. Kansas City, Missouri, USA.

Delboni, T. M., K. A. Borges, et al. (2007). "Semantic expansion of geographic web queries based on natural language positioning expressions." Transactions in GIS 3: 377-397.

Fu, G., C. B. Jones, et al. (2005). Ontology-based Spatial Query Expansion in Information Retrieval Lecture Notes in Computer Science, On the Move to Meaningful Internet Systems: ODBASE.

Goodchild, M. F. and L. L. Hill (2008). "Introduction to digital gazetteer research." Int. J. Geogr. Inf. Sci. 22(10): 1039-1044.

Gouvêa, C., S. Loh, et al. (2008). Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing. GEOINFO Rio de Janeiro, RJ.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. Proceedings of the 4th ECDL, Springer-Verlag.

Jones, C. B. and R. S. Purves (2008). "Geographical information retrieval." Int. J. Geogr. Inf. Sci. 22(3): 219-228.

Lopez-Pellicer, F. J., M. J. Silva, et al. (2010). Linkable Geographic Ontologies. GIR'10, Zurich, Switzerland.

Overell, S. E. and S. Ruger (2007). Geographic co-occurrence as a tool for gir. Proceedings of the 4th ACM Workshop on GIR. Lisbon, Portugal, ACM.

Popescu, A., G. Grefenstette, et al. (2008). Gazetiki: automatic creation of a geographical gazetteer. Proc. of the 8th ACM/IEEE-CS JCDL. Pittsburgh, PA, USA.

Rodrigues, C., M. Chaves, et al. (2006). Uma Representação Ontológica da Geografia Física de Portugal. IX Encontro de Utilizadores de Informação Geográfica, ESIG - Oeiras, Portugal.

Sanderson, M. and J. Kohler (2004). Analyzing geographic queries. GIR'04. Sheffield, UK.

Silva, M. J., B. Martins, et al. (2004). Adding Geographic Scopes to Web Resources. SIGIR'04 - Workshop on GIR, Sheffield, UK, ACM.

Smith, D. A. and G. Crane (2001). Disambiguating Geographic Names in a Historical Digital Library. Proceedings of the 5th ECDL, Springer-Verlag.

Souza, L. A., C. A. Davis-Jr, et al. (2005). The Role of Gazetteers in Geographic Knowledge Discovery on the Web. Proceedings of the 3th LA-Web Congress, IEEE Computer Society.

Souza, L. A., T. M. Delboni, et al. (2004). Locus: Um Localizador Espacial Urbano. VI GEOINFO, 22-24, Campos do Jordão, SP, Brazil.

Teitler, B. E., M. D. Lieberman, et al. (2008). NewsStand: a new view on news. Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. Irvine, California, ACM.

Toral, A. and R. Munoz (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. EACL 2006.

Uryupina, O. (2003). Semi-supervised learning of geographical gazetteers from the internet. In Proc. of the HLT-NAACL 2003 Workshop on Analysis of geographic references (1) ACL.

Volz, R., J. Kleb, et al. (2007). Towards ontology-based disambiguation of geographical identifiers. WWW2007. Banff, Canada.

Wang, C., X. Xie, et al. (2005). Web resource geographic location classification and detection. Special interest tracks and posters of the 14th in WWW. Chiba, Japan.

Wang, L., C. Wang, et al. (2005). Detecting dominant locations from search queries. Proceedings of the 28th annual international SIGIR conference on Research and Development in Information Retrieval. Salvador, Brazil, ACM.

A Geographic Annotation Service for Biodiversity Systems

Fabiana B. Gil[1,2]**, Nádia P. Kozievitch**[2]**, Ricardo da S. Torres**[2]

[1] CPqD Foundation
Rod. Campinas - Mogi-Mirim km 118.5 Campinas, SP, Brazil

[2]Institute of Computing – University of Campinas
Av. Albert Einstein, 1251 Campinas, SP, Brazil

fbellette@gmail.com, nadiapk@ic.unicamp.br, rtorres@ic.unicamp.br

Abstract. *Biodiversity studies are often based on the use of data associated with field observations. These data are usually associated with a geographic location. Most of existing biodiversity information systems provides support for storing and querying geographic data. Annotation services, in general, are not supported. This paper presents an annotation Web service to correlate biodiversity data and geographic information. We use superimposed information concepts for constructing a Web service for annotating vector geographic data. The Web service specification includes the definition of a generic API for handling annotations and the definition of a data model for storing them. The solution was validated through the implementation of a prototype for the biodiversity area considering a potential usage scenario.*

1. Introduction

The term biodiversity – or biological diversity – describes the richness and variety of biological organisms in a given habitat. Some known issues in biodiversity are irreversible loss of species, loss of environmental services (production of oxygen by plants, the hydrological balance, soil fertility, and climate balance), and biopiracy.

Computer Science can be a great allied of Biologists, providing them with tools to analyze and report findings on species and their behaviors. Some challenges for the Biodiversity Information Systems (BISs) are [Torres et al. 2006]: (i) handle large volumes of information, (ii) integrate information from different sources and formats (heterogeneity), (iii) manipulate data and images, (iv) manipulate geospatial information reference.

In this paper we address the forth challenge. Biodiversity studies are often based on the use of data associated with field observations, later matched with geographic locations. Most of existing biodiversity information systems provides support for storing and querying geographic data. Annotation services, in general, are not supported.

Annotation has been recognized as one of the most important services in digital library systems to foster the cooperation among users and the integration of heterogeneous information resources [Agosti and Ferro 2008]. In this paper, we describe a new geographic data annotation Web service that can be easily integrated with other applications. The specification and implementation of the proposed Web service relies on two main contributions: (a) the proposal of a data model based on superimposed information [Maier and Delcambre 1999a] to manage geographic annotations; and (b) the definition of a generic API to manipulate annotations.