The Role of Gazetteers in Geographic Knowledge Discovery on the Web

Ligiane A. Souza[1], Clodoveu A. Davis Jr.[2], Karla A. V. Borges[1,3], Tiago M. Delboni[1], Alberto H. F. Laender[1]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais
31270-901, Belo Horizonte, Minas Gerais, Brasil
{ligiane, kavb, delboni, laender}@dcc.ufmg.br

[2]Instituto de Informática - Pontifícia Universidade Católica de Minas Gerais
Rua Walter Ianni, 255, 31980-110, Belo Horizonte, Minas Gerais, Brasil
clodoveu@pucminas.br

[3]PRODABEL – Empresa de Informática e Informação do Município de Belo Horizonte
Av. Presidente Carlos Luz, 1275, 31230-000, Belo Horizonte, Minas Gerais, Brasil

Abstract

The Web is a large source of geographic information. Many Web documents have one or more spatial references, such as place names, addresses, zip codes or phone numbers. These spatial references are usually found in a semistructured fashion, which allows humans to identify and assign a geographic meaning to documents. In this paper, we discuss the important role that gazetteers, which are spatial catalogues of place names, can play in automating this process, and introduce the Locus gazetteer. Locus has been designed to hold not only place names for entities such as cities and rivers, but also to handle intra-urban place names, such as street names, urban landmarks, and postal addresses, along with their spatial relationships, through an ontology of places. We demonstrate that ontologically-enhanced gazetteers, such as Locus, are very useful for discovering the geographic context present on Web pages, and are often used in many other applications, such as in address geocoding for geographic information systems. To efficiently accomplish these tasks, the gazetteer must have a large database of spatial references; however, such a database is hard to obtain in emergent countries such as Brazil, in which available official geographic databases are limited and not well updated. As a way to tackle this problem, we describe a semi-automatic method used to populate the Locus gazetteer with geographic content extracted directly from the Web. To evaluate our work, an experiment was conducted, focusing on testing the Locus gazetteer data quality and comprehensiveness.

1. Introduction

Searching the World Wide Web for retrieving information based on geographic criteria is a common task in the daily routine of many people, such as tourists, who need to learn about a certain destination, citizens, who need to locate some point of interest in the city, or urban planners, who need to collect data on activities that take place in or around some region.

The Web is a huge source of this kind of information. This is why research on the Web's geographic dimension, the so-called *local Web* [11], is getting so much attention nowadays. Projects like SPIRIT[1] and Google Local[2] show that the interest on search engines able to deal with the spatial context implicit in Web documents is increasing. Studies that investigate query logs demonstrate that approximately 18% of

[1] SPIRIT Project: http://www.geo-spirit.org/

[2] Google Local: http://local.google.com

Proceedings of the Third Latin American Web Congress (LA-WEB'05)
0-7695-2471-0/05 $20.00 © 2005 IEEE

all queries contain some kind of geographic term [18]. People are looking for pages containing information about “hotels in New York City” or “restaurants near Times Square”.

However, performing such retrieval tasks over the Web is limiting in a number of ways. These limitations are partly due to the manner in which Web data are structured, but also due to the freedom that users have to specify their information needs. Besides, geographic context is often only implicit in Web pages, and cannot be trivially specified by means of keywords when we use most of the current search engines.

One of the strongest indications about the user’s intention to perform a Web search involving geographic locations is the use of place names. We then assume that the use of a catalog of place names would be an important tool to allow search engines to distinguish between regular, *subject-related* keywords, and geographic, *place-related* keywords. Such a catalog is known in the field of Geographic Information Retrieval (GIR) as a *gazetteer*. GIR is becoming increasingly important, seeking solutions to problems such as the identification of geographic context in plain text [20], resolution of ambiguities in place names [16], building spatial indexes for documents [6] and others.

We realize that the usual definitions for a gazetteer need to be extended somewhat to accommodate the needs of GIR. First, we must consider ways to organize gazetteer data in order to facilitate retrieval by search engines; this requires novel ways to model and develop the gazetteer, going well beyond its original uses in geography and cartography [10]. Additionally, a rich and continuously renewable source of data to populate the gazetteer as automatically as possible is required, in order to maximize its efficiency in recognizing relevant place names. In this paper, we will show our response to both of these challenges. We present Locus, a gazetteer that has been designed to hold not only place names for entities such as cities, rivers, mountains, and so on, but also to handle intra-urban place names, such as street names, urban landmarks and postal addresses along with spatial relationships among them, defined through an ontology of places [19]. We also present a Web data collection strategy and tool to populate Locus with place names extracted from the text in selected Web sites.

This paper is organized as follows. Section 2 presents an introduction on geographic information retrieval on the Web. Section 3 presents concepts on digital gazetteers and explains our view on gazetteers being essential assets to explore the Web’s geographical dimension. In section 4, we present Locus, a spatial locator system based on a gazetteer and developed by the UFMG Database Group. Section 5 describes the process used to extract references from the Web to populate the Locus gazetteer. Section 6 describes the Locus user interface. In Section 7, we present results from an experiment conducted for evaluating the Locus gazetteer data quality and comprehensiveness. Finally, section 8 presents our conclusions.

2. Geographic information retrieval on the Web

Larson [15] defines *geographic information retrieval* as an applied research area that combines aspects of databases, human-computer interaction, geographic information systems, and information retrieval. GIR is concerned with indexing, searching, retrieving, and browsing georeferenced information sources, and with the design of systems to accomplish these tasks effectively and efficiently.

In GIR, exploring geographic aspects of the Web can take place according to two different approaches [17]. The first (*entity-based*) approach uses Internet infrastructure elements to obtain information about the physical location of hosts. Exploring the Web in this way allows content to be deployed considering the user’s inferred location, making online advertising much more effective. Recently, the Stanford Research Institute (SRI) formulated a proposal to create a top level domain *.geo* “to provide a complete, virtually free-of-charge, and open infrastructure for referencing and discovering georeferenced information on the Internet” [12]. The concept is based on a grid representation of the Earth’s surface that would be used to associate a location to servers and other entities in the network. As a consequence, each conventional URL could be automatically translated to an approximate geographic coordinate. Disregarding for the moment the SRI proposal, locating users based on their server’s inferred location can be quite imprecise, even though results can be obtained quickly.

The second approach is called *content-based*, since it uses elements contained in the pages themselves to deduce a location, or a number of locations, referred to by the page. A Web page can contain indications of its connection to a specific place embedded in text, in the form of place names, postal addresses, ZIP codes, phone numbers and area codes, and so on. The challenge here involves the retrieval, semantic analysis and interpretation of the indications, leading to the connection of the Web page to a number of geographic locations. Even though this approach is much more complex than the previous one, we observe that a common problem with the previous approach is avoided: the fact that pages referring to a location can be stored in servers located elsewhere.

Information retrieval methods used in search engines are typically limited to keyword searches and link analysis, offering no support for spatial context exploration in documents. Realizing this limitation, search engine developers are studying new methods to improve the quality of results to their users' spatial queries. Google Local uses a yellow pages directory covering the United States, Canada and U.K. to spatially index documents containing references to services such as shops, hotels, and restaurants. SPIRIT (Spatially-aware Information Retrieval on the Internet), a search engine still under development, plans to go beyond that, using geographic ontologies to produce spatial metadata from Web pages. Based on the knowledge contained in the ontologies, each geographic term found in the page is extracted and linked with a spatial footprint. After that, footprints associated with the page are used to build a spatial index for the search engine.

3. Digital gazetteers

Gazetteers are *geospatial dictionaries of geographic names* [10]. A gazetteer is a powerful tool that makes it possible to identify *indirect geographic references* (place names) on documents and to associate geographic coordinates to them.

Three attributes are essential for any record in a gazetteer: the *name* of the place, its *footprint* (spatial location), and its *type* [10]. With these attributes, a gazetteer is able to answer at least two basic queries: to find a place given its name, and to find names associated to a given place.

Many gazetteers are currently available on the Web. One of the most important is the Alexandria Digital Library Gazetteer (ADL)[3]. With approximately 4.4 million entries, ADL's gazetteer data storage is based on a comprehensive metadata structure, called Gazetteer Content Standard. ADL introduced a significant innovation: a Web service protocol to allow free access to the gazetteer. The service requests and answers are XML documents following the Gazetteer Content Standard structure.

According to Fu et al. [6], current gazetteers share a number of limitations that keep them from being used in GIR. First, most gazetteers do not encode spatial relationships, apart from simple region hierarchies. Second, generic relationships between object types are not implemented, therefore limiting the potential use of the gazetteer as a geographic ontology. Furthermore, properties of geographic features are only generically defined, and usually lack significant details. Third, gazetteers usually contain only names associated to well-defined footprints, and lack support for fuzzy or imprecise locations, such as "southern California". To the observations in [6], we add the lack of intra-urban names, such as city landmarks, monuments, and other well-known locations used by citizens as reference points for navigation and to indicate the location of other points of their interest in a local context.

We believe that the development of new digital gazetteers, designed to maintain urban data and to process queries involving spatial relationships, can greatly expand their usefulness in local Web exploration.

4. The Locus system

Locus, a spatial locator system, was conceived as a tool to support the recognition of spatial context on Web pages (see [19] for an early design of Locus). The Locus design is based on a formally specified ontology, the *ontology of places*.

In the traditional systems modeling approach, the modeler is required to capture a user's view of the real world in a formal conceptual schema. In doing so, the modeler follows an established paradigm, such as the entity-relationship model [2], forcing the mapping of concepts acquired from the real world to instances of abstractions available in the chosen paradigm. This process

[3] Alexandria Digital Library Gazetteer: http://www.alexandria.ucsb.edu/gazetteer/

tends to introduce inconsistencies and inaccuracies in the system development [5]. An ontology is an explicit specification of a conceptualization [8]. Ontologies are semantically richer than conceptual schemas, and thus closer to the user's cognitive model. Guarino coined the term *ontology-driven information systems* for systems that make use of formally defined ontologies [9].

The ontology of places defines concepts from the particular domain of urban geographic space. The hierarchy of territories, in which regions are subdivided into other regions, is explored by the ontology, as well as concepts related to urban addresses and landmarks commonly used by the population. Telephone numbers and postal codes, which can be used as indirect location identifiers, are also considered by the ontology. The ontology's spatial knowledge elements can be used to infer geographic relationships between objects, such as proximity, adjacency, containment and so on.

Based on this ontology, a conceptual schema was specified, using the OMT-G model [1]. The main part of the schema, demonstrating the three basic classes used in the gazetteer (territory, address and landmark) and their relationships, is presented in Figure 1. The Gazetteer data is stored using the PostgreSQL DBMS, with the PostGIS module, which adds spatial storage support.

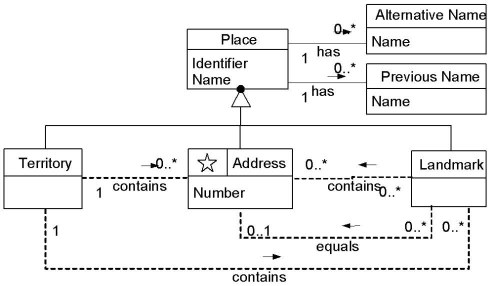


Figure 1 – Gazetteer conceptual schema

Figure 2 describes the Locus architecture. The query processor module is the main Locus component. In Locus, all queries follow the canonical form for geospatial data queries, as proposed by Egenhofer [3]. Two types of queries are available: simple and advanced. Simple queries contain only geospatial terms, which can be the name of a place or a type of place plus a name. Advanced queries are defined as a triple (P, S, L), where P is the type of a point of interest that is being sought by the user, *L* is the name of a point of reference, a well-known landmark that serves as a rough geographic location for the query, and *S* is a natural-language expression that indicates the expected spatial relationship between P and L. This allows the construction of elaborate queries, such as "restaurant" "near" "Times Square", or "subway station" "in" "Financial District".

Spatial relationships can be topological (inside of), metric (500 meters from), fuzzy (near) or ordered (in front of) [4]. Fuzzy relationships are converted by Locus to metric relationships, based on a heuristic which considers the dimensions of the objects involved in a query. If the involved place has a point representation, all other places in a radius of 500 meters are considered to be near. If the place is represented as a line or polygon, its minimum bounding rectangle is expanded by about 20% to determine the places considered to be "near".

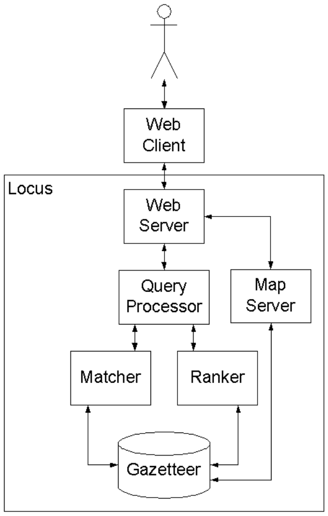


Figure 2 – Locus architecture

During the processing, a query first passes through the matcher module. The matching function finds all place names that match the query terms exactly or approximately, i.e., allowing a limited number of errors in the strings. To compute the approximate string matching, the Shift-Or algorithm [21] is used. In advanced query processing, after finding the landmark, a perimeter for the location of the point of interest is established, based on the semantics of the specified spatial relationship. For instance, in the query "restaurants" "near" "Times Square", first we locate Times Square, then apply a buffer around its location, corresponding to the notion of "near" in an urban context (since Times Square is catalogued by the gazetteer as an urban object). Finally, restaurants contained in that area are selected and displayed.

Many places share the same name. This forces the matcher module to return many responses for some queries. In these cases, the *ranker* module is executed to sort the answer, in order to show the most significant places in the first positions. This classification is based on an index, which is calculated considering the number of errors in the string matching and an importance factor, assigned to the places. The parameter used to calculate the relative importance for territories and streets is their population, when this data is available. We assume that a larger population implies a greater likelihood that the place is the one being sought by the user. Landmark ranking, i.e., an automated method to infer the importance of reference places, is still a problem, but techniques based on the number of times the landmark name is mentioned on the Web are being studied.

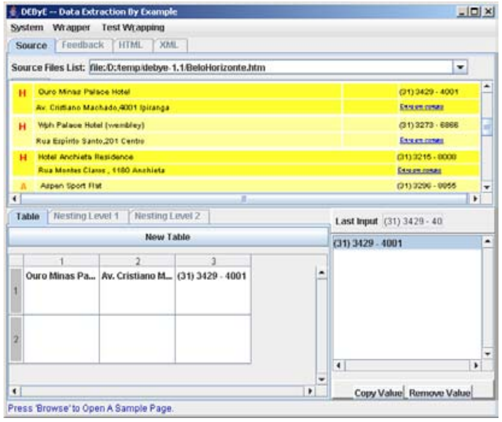
5. Populating the Locus gazetteer

Our current version of the Locus gazetteer has been populated with data from Brazilian cities. However, one problem we found during the implementation was the lack of reliable sources of data. At first, we collected public records from organizations such as IBGE (the Brazilian mapping agency), the Brazilian postal company, and some local government institutions. From these sources, the Locus gazetteer was initially populated with approximately one million Brazilian places, including urban places from the city of Belo Horizonte. These sources are very reliable and thorough, but they are limited to some kinds of official place names, such as city, district, and neighborhood names. We also needed data on possible points of interest, especially on services such as hotels, restaurants, commercial centers, and others, and on places used as urban landmarks by the population, such as monuments, buildings, squares, and many others. Such data were available for the city of Belo Horizonte, but were very hard to obtain for other Brazilian cities, in which GIS development is at its earlier stages.

As proposed by Laender et al. [14], the strategy used to expand the gazetteer database was to collect pages from Web sites containing tourist information, and to extract references to services from these pages, including their names, types, telephone numbers and addresses.

Data extraction from the Web, which actually includes page collection, data identification, and data extraction from collected pages, is executed by programs called *wrappers*. We used the WByE (Wrapping By Example) environment for this purpose [7]. Also developed by the UFMG Database Group, WByE is able to create wrappers based on examples supplied by the user. From these examples, specifications are generated to build agents responsible for page collection and data extraction.

WByE comprises two components: ASByE (Agent Specification By Example) and DEByE (Data Extraction By Example). ASByE is responsible to generate a *page fetching plan* (PFP), used for the collection of pages. DEByE is used to determine how data appear in the pages and how the extraction result should be logically organized. To accomplish this, based on user supplied examples, DEByE generates an *object extraction pattern* (OEP) that captures the target object's structure and specifies how data should be extracted accordingly [13]. Figure 3 shows the DEByE interface, where examples for data values to be extracted from a page containing information on hotels are specified. These examples are taken from a sample page and pasted into the columns of the table located at the left hand side of the bottom part of the screen.



Source Files List: file:D:/temp/debye-1.1/BeloHorizonte.htm

H Ouro Minas Palace Hotel (31) 3429 - 4001
Av. Cristiano Machado,4001 Ipiranga
H Wph Palace Hotel (wembley) (31) 3273 - 6866
Rua Espirito Santo,201 Centro
H Hotel Anchieta Residence (31) 3215 - 0000
Rua Montes Claros , 1180 Anchieta
A Aspen Sport Flat (31) 3296 - 0056

Last Input (31) 3429 - 40

(31) 3429 - 4001

	1	2	3
1	Ouro Minas Pa...	Av. Cristiano M...	(31) 3429 - 4001
2			

Press 'Browse' to Open A Sample Page.

Figure 3 – DEByE example specification

The complete process is presented in Figure 4. First, users supply examples to ASByE on how to collect the pages from the selected Web sites. The resulting PFP is used to generate a fetching module, which effectively collects the pages from the Web. The user intervenes again in the process by supplying examples to DEByE of how to extract the references from selected pages. An extracting module is generated from the respective OEP. The references are extracted and stored in a XML repository. To avoid reference repetitions, the Similarity Analyzer Module evaluates the references and inserts in the gazetteer only those that were not yet stored. The insertion includes georeferencing based on the available information about the reference. If an address was extracted, a geocoding process is executed to associate the reference with its individual address in the gazetteer (if it is available). If there is no address information, the georeferencing occurs in a less precise manner, generating a pointer to the corresponding neighborhood or municipality.

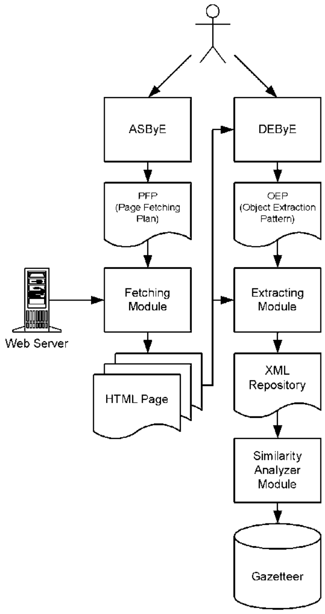


Figure 4 – Gazetteer population process

In a preliminary experiment to evaluate the effectiveness of this strategy, we extracted references from six selected Brazilian Web sites. Six reference attributes were extracted: *name*, *type*, *address*, *city*, *state* and *telephone*. Name, city and state were considered mandatory and references without any of these attributes were not considered.

The similarity analyzer procedure is described in Figure 5 . Each candidate reference to insertion is compared with all references already stored in the gazetteer using this procedure. The insertion occurs only if the result is false for all the tests.

```
TestSimilarity (r1, r2)
begin
  if r1(city, state) != r2(city, state)
    return false;
  if  r1(telephone) == r2(telephone)
    return true;
  if r1(address) == r2(address)
    if EditDistance  (r1(name),  ref2(name))  <
given limit
      return true;
  if EditDistance (r1(name), r2(name)) < given
limit
    if r1(type) == r2(type)
      return true;
  return false;
end;
```

Figure 5 – Similarity analyzer algorithm

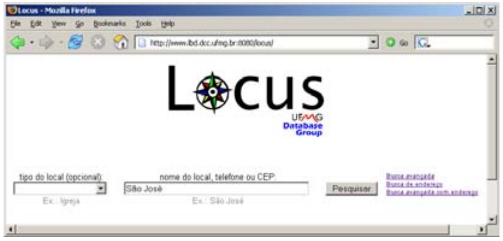
Table 1 shows the selected Web sites, the number of pages collected from each one and the number of references extracted with this method. After the execution of the similarity analysis, 29,139 references were effectively inserted in the gazetteer. Additionally, 1,715 references that were already stored were improved, by the insertion of an address or a telephone number.

Web site	Collected pages	Extracted references
cidades.terra.com.br	278	2,443
www.cidades.com.br	1,416	10,588
www.guiadasemana.com.br	15	2,977
www.citybrazil.com.br	4,900	4,628
www.ondehospedar.com.br	1,131	12,286
www.pachecodrogaria.com.br	1	393
Total	7,741	33,315

Table 1 – Data extraction results

6. The Locus user interface

Figure 6 shows the Locus simple query interface, which can be accessed at the URL http://www.lbd.dcc.ufmg.br:8080/locus. Using this interface, a user formulates a query by specifying a place name (or a place locating information such as zip code or phone number) and, optionally, a place type (e.g., city, street, church, etc.).



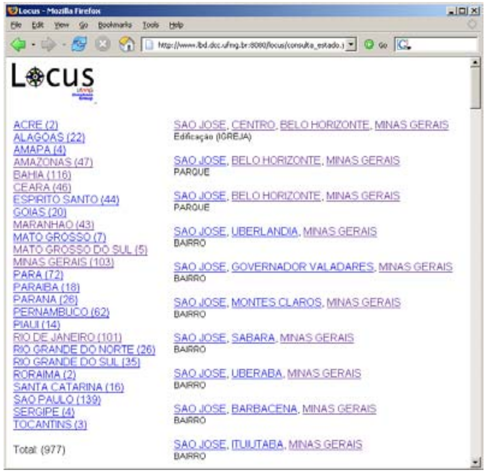
http://www.lbd.dcc.ufmg.br:8080/locus/

tipo do local (opcional):

nome do local, telefone ou CEP: São José

Figure 6 – Simple query interface

If more than one place is found for a query, a list containing type, name and location of all found places is presented for user selection. Figure 7 shows the list of places found for the query “São José” (“Saint Joseph”). The first place returned, for instance, is the Saint Joseph Church in the city of Belo Horizonte.


Locus - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.lbd.dcc.ufmg.br:8080/locus/consulta_estado.j
Go
Locus
UFMG Database Group

ACRE (2)
ALAGOAS (22)
AMAPA (4)
AMAZONAS (47)
BAHIA (116)
CEARA (46)
ESPIRITO SANTO (44)
GOIAS (20)
MARANHAO (43)
MATO GROSSO (7)
MATO GROSSO DO SUL (5)
MINAS GERAIS (103)
PARA (72)
PARAIBA (18)
PARANA (26)
PERNAMBUCO (62)
PIAUI (14)
RIO DE JANEIRO (101)
RIO GRANDE DO NORTE (26)
RIO GRANDE DO SUL (35)
RORAIMA (2)
SANTA CATARINA (16)
SAO PAULO (139)
SERGIPE (4)
TOCANTINS (3)

Total: (977)

SAO JOSE, CENTRO, BELO HORIZONTE, MINAS GERAIS
Edificação (IGREJA)

SAO JOSE, BELO HORIZONTE, MINAS GERAIS
PARQUE

SAO JOSE, BELO HORIZONTE, MINAS GERAIS
PARQUE

SAO JOSE, UBERLANDIA, MINAS GERAIS
BAIRRO

SAO JOSE, GOVERNADOR VALADARES, MINAS GERAIS
BAIRRO

SAO JOSE, MONTES CLAROS, MINAS GERAIS
BAIRRO

SAO JOSE, SABARA, MINAS GERAIS
BAIRRO

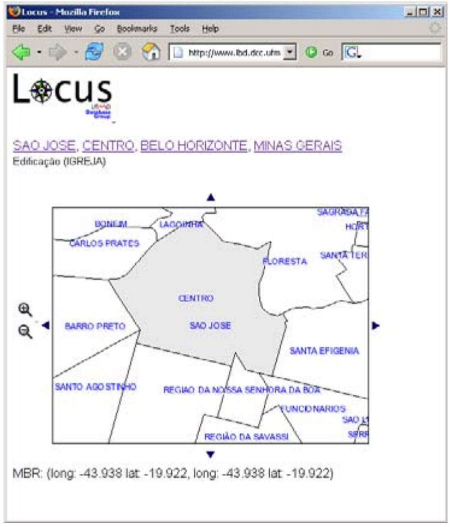
SAO JOSE, UBERABA, MINAS GERAIS
BAIRRO

SAO JOSE, BARBACENA, MINAS GERAIS
BAIRRO

SAO JOSE, ITUIUTABA, MINAS GERAIS
BAIRRO

Figure 7 – Selection interface

The answer interface appears in Figure 8. A map of the located area is shown, along with all information available about the place.

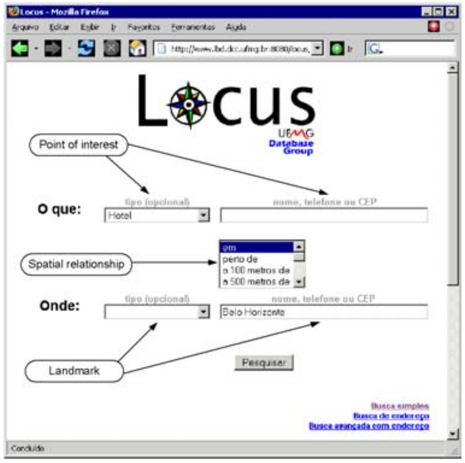


SAO JOSE, CENTRO, BELO HORIZONTE, MINAS GERAIS
Edificação (IGREJA)

MBR: (long: -43.938 lat: -19.922, long: -43.938 lat: -19.922)

Figure 8 – Answer interface

In the advanced query interface (Figure 9), the point of interest, the spatial relationship and the landmark must be informed. As for a simple query, if more than one place is found as a result, a list of options is presented to the user.



O que: tipo (opcional): Hotel; nome, telefone ou CEP:

em
perto de
a 100 metros de
a 500 metros da

Onde: tipo (opcional): ; nome, telefone ou CEP: Belo Horizonte

Figure 9 – Advanced query interface

7. Experimental evaluation

An experiment was conducted to evaluate our work, focusing on testing the Locus gazetteer data quality and comprehensiveness. Our final goal with the experiment was to estimate the gazetteer’s ability to serve as a filter for the processing of spatial queries in the context of a search engine.

Google was the selected search engine for the experiment. We know Google is not yet tuned to support geographic queries and, thus, this comparison is not completely adequate. Our purpose is just to evaluate if a gazetteer would help search engines to improve their results.

For the experiment, we first selected queries containing spatial terms that have been submitted to the TodoBR Brazilian search engine[4]. Six months of log queries were analyzed. The terms used to select spatial queries were place names, place types, spatial relations and adjectives indicating locality (e.g., mineiro, paulista). Of the total of 1.4 million queries that TodoBR received in six months, 14.1% contained one or more of these geographic terms, a figure closed to that reported by Sanderson and Kohler [18].

Next, 70 spatial queries were randomly chosen and distributed to 18 users, which were asked to submit each query to Google and evaluate the first 25 returned answers. By doing so, we could estimate Google’s success rate in finding geographic context in the queries. After

[4] TodoBR: http://www.todobr.com.br. This search engine was recently acquired by Google.

this, users would try to find the same places using Locus. Only the first 25 references returned by Locus where considered.

The graphic in Figure 10 shows users evaluations about Google's responses. The search engine has returned relevant documents for 45 geographic queries, which is equivalent to 64%. Users felt partially satisfied for 13 queries (19%) and absolutely not satisfied for 12 queries (17%).

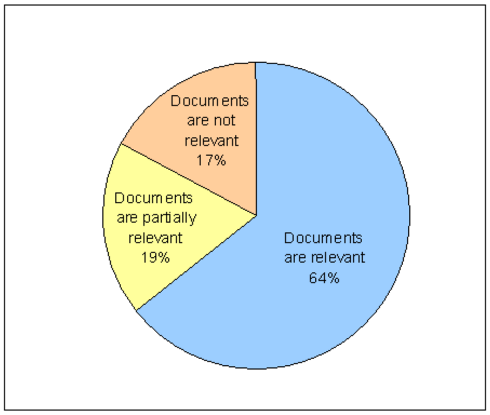


Figure 10 – Google's relevance for spatial queries

Figure 11 shows the success Locus has demonstrated on finding the places contained in the same queries had applied to Google. Locus has successfully identified 54 places (77%).

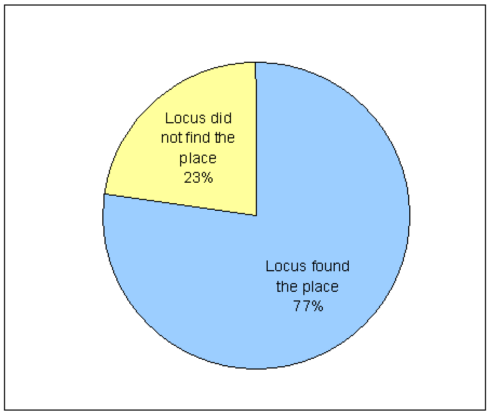


Figure 11 – Locus gazetteer success on finding places in queries

Locus was able to positively identify more places in the queries (77%) than the search engine was able to retrieve relevant documents about the places (64%). Since Goggle does not give a geographic interpretation to the queries, it throws away valuable information, which makes the search engine less efficient on responding to user expectations. We strongly believe gazetteers are a natural way to address this limitation.

8. Conclusions

Geographic information retrieval is in its early stages. The discovering of geographic context on Web documents is not a trivial task and the interest in GIR tends to expand greatly in the next years.

In this paper, we have discussed the important role gazetteer can take in this scenario. Our intended contributions were to present Locus, a spatial locator system based on a digital gazetteer, and the semi-automated method we used to expand the Locus gazetteer database. In an experiment, we have evaluated Locus' ability to help in the identification of geographic context in search engines queries.

In spite of the limitations in Locus, we have found it to be already useful for its purpose. Future works include the development of a better ranking procedure and improvements on the user interface. We are also considering the development of a Web service based on Locus, so that it can be integrated to information systems and act as a support for the discovery of geographic context in natural language texts.

Acknowledgments

This work was partially funded by the GERINDO Project (MCT/CNPq/CT-INFO). Karla A. V. Borges acknowledges the institutional support from Prodabel for her doctoral program.

References

[1] Borges, K. A. V., Davis Jr., C. A., Laender, A. H. F. OMT-G: An object-oriented data model for geographic applications. GeoInformatica, 5(3): 221-260, 2001.

[2] Chen, P. P. The Entity-Relationship Model - Toward a Unified View of Data. ACM Transactions on Database Systems. 1(1): 9-36, 1976.

[3] Egenhofer, M. J. Toward the semantic geospatial web. Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, McLean, Virginia, pp. 1-4, 2002.

[4] Egenhofer, M. J., Franzosa, R. D. Point-set topological spatial relations. International Journal of Geographical Information Systems,

5(2): 161-174, 1991.

[5] Fonseca, F. T., Davis, C. A., Camara, G. Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. GeoInformatica, 7(4): 355-378, 2003

[6] Fu, G., Abdelmoty, A. I., Jones, C. B. Design of a Geographical Ontology. Available at www.geo-spirit.org/publications/spirit_wp3_d5.pdf. Last access on May 2005.

[7] Golgher, P. B., Laender, A. H. F., Silva, A. S., Ribeiro-Neto, B. A. An Example-Based Environment for Wrapper Generation. Proceedings of the Workshops on Conceptual Modeling Approaches for E-Business and The World Wide Web and Conceptual Modeling: Conceptual Modeling for E-Business and the Web, Lecture Notes in Computer Science, 1921: 152-164, 2000

[8] Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies, 43 (5/6), pp. 907-928, 1995.

[9] Guarino, N. Formal Ontology and Information Systems. Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, Trento, Italy, pp. 3-15, 1998.

[10] Hill, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, 1923, pp. 280-290, 2000.

[11] Himmelstein, M. Local Search: The Internet Is the Yellow Pages. IEEE Computer 38(2): 26-34. 2005.

[12] ICANN - The Internet Corporation for Assigned Names and Numbers. Summary Application of SRI International. Available at http://www.icann.org/tlds/report/geo1.html. Last access on May 2005.

[13] Laender H. F. A., Ribeiro-Neto, B. A., Silva, A. S. DEByE - Data Extraction By Example. Data & Knowledge Engeneering. 40(2): 121-154, 2002.

[14] Laender, A. H. F., Borges, K. A. V., Carvalho, J. C. P. ; Medeiros, C. B., Silva, A. S., Davis Jr., C. A. . Integrating Web Data and Geographic Knowledge Into Spatial Databases. In: Yannis Manolopoulos; Apostolos Papadopoulos; Michael Vassilakopoulos. (Org.). Spatial Databases: Technologies, Techniques and Trends. Hershey, Pennsylvania, USA, pp. 23-48, 2004.

[15] Larson, R. R. Geographic Information Retrieval and Spatial Browsing. GIS and Libraries: Patrons, Maps and Spatial Information, University of Illinois, pp. 81-124, 1996.

[16] Leidner, J. L. Toponym Resolution in Text: “Which Sheffield is it?”, Proceedings of the 27th Annual International Conference on Research

[17] McCurley, K. S., Geospatial Mapping and Navigation of the Web, Proceedings of World Wide Web International Conference, Hong Kong, pp. 221-229, 2001.

[18] Sanderson, M., Kohler, J. Analyzing Geographic Queries. ACM SIGIR 2004 Workshop on Geographic Information Retrieval. Available at http://www.geo.unizh.ch/~rsp/gir/abstracts/sanderson.pdf. Last access on May 2005.

[19] Souza, L. A., Delboni, T. M., Borges K. A. V., Davis Jr., C. A., Laender A. H. F. Locus: Um Localizador Espacial Urbano, Proceedings of GeoInfo 2004 - VI Brazilian Symposium on Geoinformatics, Campos do Jordão, pp. 467-478, 2004.

[20] Woodruff, A. G., Plaunt, C. GIPSY: Geo-referenced Information Processing System. Journal of the American Society for Information Science, 45: 645-655, 1994.

[21] Wu, S., Manber, U. Fast Text Searching Allowing Errors. Communications of The ACM, 35(10): 83-91, 1992.