Discovering Geographic Locations in Web Pages Using Urban Addresses

Karla A. V. Borges
Prodabel

Av. Carlos Luz, 1275
31230-000 Belo Horizonte
MG Brazil
karla@pbh.gov.br

Alberto H. F. Laender
Federal University of
Minas Gerais
Av. Antônio Carlos, 6627
31270-010 Belo Horizonte
MG Brazil
laender@dcc.ufmg.br

Claudia B. Medeiros
State University of
Campinas
Caixa postal 6176
13084-971 Campinas SP
Brazil
cmbm@ic.unicamp.br

Clodoveu A. Davis Jr.
Pontifical Catholic University
of Minas Gerais
R. Walter Ianni, 255
31980-000 Belo Horizonte
MG Brazil
clodoveu@pucminas.br

ABSTRACT

This paper presents an approach that helps to discover geographic locations from the recognition, extraction, and geocoding of urban addresses found in Web pages. Experiments that evaluate the presence and incidence of urban addresses in Web pages are described. Experimental results, based on a collection of over 4 million documents from the Brazilian Web, show the feasibility and effectiveness of the proposed method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Language Constructs and Features – *query formulation, retrieval models, search process*.

General Terms

Design, Experimentation, Performance.

Keywords

Geographic information retrieval, local search, urban ontology, urban address.

1. INTRODUCTION

Automatic recognition of geographic characteristics embedded in Web data and documents has countless social and economic applications, e.g., in tourism or health. However, it remains a difficult task. Many Web pages contain geospatial evidence such as place names, addresses, postal codes, or phone numbers, usually in a semi-structured fashion, nevertheless allowing humans to recognize it and assign geographic meaning to Web pages. Other evidence is found embedded in natural-language text, and recognizing it requires understanding the page's context.

Sanderson and Kohler [14] verified that about 18% of keywords submitted as queries to the Excite search engine contain geography-related terms. In Brazil, an analysis of six months of query logs from TodoBR (a major Brazilian search engine, acquired by Google in 2005) [6] revealed that 14.1% of the queries contained at least one geographic-related term such as a place name or type, a spatial relation, or an adjective indicating locality. Moreover, at least 20% of the Web pages included one or more easily recognizable, unambiguous geographic identifiers, such as postal addresses. These pages usually include locally relevant content [1, 3, 6, 9, 13].

These numbers show that people are looking for Web pages containing faster, more useful information about everyday tasks: local merchants, services, and news are frequently sought [9]. However, traditional search engines only recognize limited geographic context in Web pages and produce results that are not geographically significant. Because of that, some search engines recently introduced mapping and routing capabilities as a doorway into local search. To achieve that, they must keep a sort of yellow pages directory on the Web, and then add functions to locate businesses on maps or on high-resolution satellite imagery.

While using geographic criteria in Web searches is increasingly common, recognizing such geographic references remains a challenge. It involves retrieval, semantic analysis, and interpretation of the geography-related evidence in each Web page before determining or approximating the correct location. At least three steps are necessary. First, distinguish place names from other words in natural-language text with little or no structuring. Second, find ways to filter coincidental names that refer to several different places. For example, "Savoy" can refer to a region in France, a hotel in London, or a restaurant in Vienna. Third, determine the correct location. Therefore, a new approach which recognizes geographic references and understands their contexts in Web pages will significantly improve local search accuracy.

We consider that next-generation Web mapping tools will integrate Web pages and maps with better local search facilities. These facilities should combine the term-based approach employed by search engines with urban locations, using the elements contained in the pages. In this context, it is important to develop new methods for recognizing local geographic references within Web pages, and understanding the context in which these references are used, so we can extract geographic knowledge from them. This paper presents an ontology-based approach that helps recognize, extract, and geocode geospatial evidence with local characteristics, such as street names, urban landmarks, telephone area codes, and postal addresses. Our focus is on extracting geographic knowledge from business or local service pages, which are the ones that provide more useful local information.

GIR'07, November 9, 2007, Lisbon, Portugal.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 presents our proposal, showing how geospatial evidence with local characteristics can be recognized and extracted from Web pages. Section 4 shows results obtained from an experiment with Brazilian Web pages. Section 5 presents conclusions and directions for future work.

2. RELATED WORK

Previous works [1, 4, 8, 10, 13, 15, 18, 19] have already considered using the intended meaning of terms, expressions and phrases in natural language as a useful paradigm for navigating and retrieving geographic information from the Web. Larson [12] defines Geographic Information Retrieval (GIR) as an applied research area that combines aspects of databases, human-computer interaction, Geographic Information Systems (GIS), and Information Retrieval (IR). GIR is concerned with indexing, searching, retrieving, and browsing georeferenced information sources, and with the design of systems to carry out these tasks effectively and efficiently. In GIR, exploring geographic aspects of the Web can take place according to two different approaches [1]. The first approach (*Source Geography*) uses Internet infrastructure elements to obtain information about the physical location of hosts. This allows content to be deployed considering the user's inferred location, making online advertising more effective. The second approach (*Target Geography*) uses elements contained in the page to deduce a location, or several locations, to which it refers. Such elements include place names, postal addresses, ZIP codes, phone numbers and area codes, and so on. The challenge here involves the extraction, semantic analysis and interpretation of the indications, leading to the connection of the Web page to geographic locations. With the second approach, a common problem with Source Geography is avoided: pages referring to a location can be stored in servers located elsewhere.

Borges at al. [4] describe a user-assisted environment that allows the extraction of geospatial data from Web pages, converts them to XML format, and uploads the converted data into spatial databases for later use in urban GIS. Fu et al. [8] and Silva et al. [15] use geographic ontologies to obtain spatial metadata from Web pages. Based on knowledge from the ontologies, each geographic term found in the page is extracted and linked with a spatial footprint. Footprints associated with the page are then used to build a spatial index for the search engine. Embley [7] presents an approach for extracting and structuring information from data-rich unstructured documents using extraction ontologies. "Data-rich" suggests documents that have several identifiable elements, such as date, name, and time.

Some commercial search tools have recently started to offer geographic search capabilities, allowing the user to locate places of interest near a given address and to navigate on the selected Web sites. Services such as Google Local use yellow pages business directories to retrieve information associated with locations within a given distance of a specified search center. Himmelstein [9] discusses the rapid growth of *local search*, a kind of geographically oriented search, and explains why this subject is attractive to both the commercial and research sectors. On-line local search uses addresses for efficient proximity estimation.

Our approach differs from the ones just mentioned, since we focus on the local Web – i.e., pages concerning a given region – and on pages corresponding to an urban location. It is based on an extraction ontology of urban places that helps recognizing, extracting, and geocoding complete or partial urban addresses.

3. GEOSPATIAL EVIDENCE IN WEB PAGES

Recognition of geographic context is a complex task, since text with geographic meaning can occur anywhere on a page. When the evidence is a place name, the problem becomes even more complicated, since there can be ambiguities. Most works found in the literature use place names as the main geospatial evidence within a page. Alternative intraurban evidence, such as address, phone number, and postal code, are much less explored.

The urban address is, among all types of urban geospatial evidence, the most adequate for local search applications, since it is closely associated to an urban place and represents the physical location of services and activities included in Web pages. Even though the recognition of a postal address is a well-studied problem, especially in GIS, when dealing with Web pages the lack of a universal standard complicates things [13]. Address formats vary widely among countries, and variations on elements such as abbreviations, punctuation, line breaks and others make the development of an address parser a nontrivial task. On the other hand, within a given country, the recognition of postal addresses embedded in natural language text can be well established.

Web page authors often leave the country name implicit in postal addresses [13]. In those cases, the use of other pieces of evidence is essential to determine the location. Postal codes are a naturally strong evidence of location, since their recognition allows for the direct association of the page to a specific part of a country. Ground line phone numbers implicitly carry information on location, since numbering is organized according to geographic principles, in order to provide efficient cabling and equipment distribution. Area and prefix codes identify the country and the city with minimal ambiguity. Recognizing phone numbers in Web pages requires some precautions, so as not to confuse them with other data, such as serial numbers. Since most traffic is local, often phone numbers omit area codes, and do not include country codes. There is also a wide variation on the use of separators, such as dashes, parentheses, and blanks. A parser for phone numbers must be flexible enough to accommodate such variations.

4. ADDRESS RECOGNITION

There are several different strategies for extracting data from Web pages [11]. Ours is based on extraction ontologies [7] that define standards and rules designed for the recognition and extraction of data of interest. Therefore, this section discusses recognition strategies and standards for the extraction of geospatial evidences, according to the definition of address established by an ontology on urban places, called OnLocus, previously developed by the authors [3] (partially shown in Figure 1). OnLocus has been proposed as a semantic support for the recognition, interpretation and extraction of terms that refer to urban places.

According to definitions contained in OnLocus, an address can be divided into three parts (Figure 2). The *basic address* supplies a street type, its name, and a building number. The second part is optional, and provides a *complement* to the basic address, including neighborhood name. The third part, called *location*, is subdivided into three identifiers used to locate an address in a definite urban context: postal code, phone number, and city/state. For the recognition of an address, only the first part is required. However, if one wants to know its location, at least one of the three location identifiers must be present.

An address can be found in a complete, incomplete, or partial form. In the complete form, the address contains the basic part, with all location identifiers. An incomplete address contains the basic part, plus at least one of the location identifiers. A partial address only includes location identifiers. Language and local culture must be observed in address recognition. Even though addresses are used worldwide and are formed of essentially the same components, the sequence in which these components appear varies among countries. The parser must be able to deal with the order of address components for the extraction.

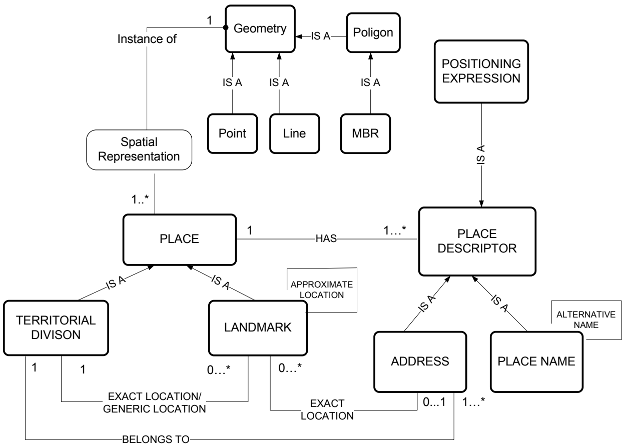

Figure 1- Partial graphical representation of OnLocus

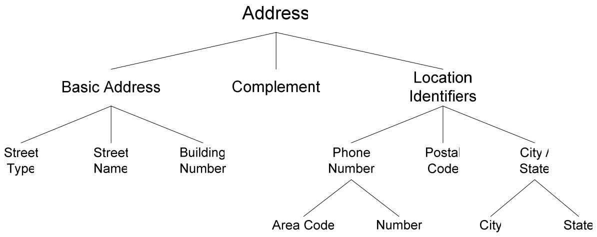

Figure 2 – OnLocus – Address structure

4.1 Address Extraction

Figure 3 presents the main steps of the proposed process for recognition, extraction and geocoding of addresses from Web pages. Initially, Web pages are collected and preprocessed (1). Preprocessing selects HTML documents and normalizes the set of tokens and delimiters, reducing the complexity for recognition patterns. During preprocessing, duplicate pages and non-HTML documents are discarded. The remaining pages have their tags replaced by a special marker, and are stripped of accent marks, control characters, and consecutive spaces. Pages then move on to the recognition and extraction of potential geospatial references, such as addresses, postal codes, and phone numbers (2) using geoparsing. Geoparsing only recognizes addresses that include at least one of the location identifiers, considering that every address at this step is potentially in a city. Geoparsing results are structured addresses, extracted and forwarded to the geocoding step (3). Geocoding results are then stored in a repository.

Geocoding is the process that determines coordinates based on alphanumeric data [5]. Before it takes place, extracted elements such as postal code, telephone area code or city/state names are checked against a gazetteer. If there is a match, the location is recognized and validated, and geocoding proceeds in two stages: *matching* and *locating*. In the matching stage, a correspondence is established between the identified address and a geographic entity from the gazetteer (such as a street, neighborhood, or city). In the locating stage, geographic coordinates are associated with the address. Geocoding considers the existence of an addressing infrastructure, including point-georeferenced individual addresses and street segments associated to numbering ranges. Results can be *exact*, when the extracted address corresponds exactly to an address that is available in the gazetteer, or *approximate*, when the location is estimated from nearby elements, such as the closest building number within the street or the street segment that includes the provided building number. If neither an exact nor an approximate location can be determined, a *generic location* can often be established, using information such as the neighborhood name, postal code, or city limits [5].

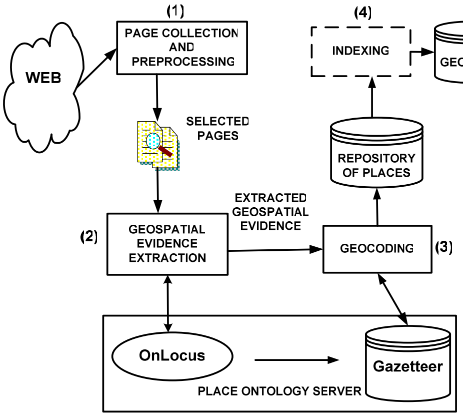

Figure 3 – Recognition, extraction and geocoding of geospatial evidences from Web pages

As a result of the geocoding, a *repository of places* is formed. This repository contains, for each extracted and validated address, the URL in which it was found, the pattern used in the extraction, the extracted terms, the initial and final position of the terms in the page, the city and state names, the minimum bounding rectangle (MBR) of the city's limits, and, if possible, the geographic coordinates associated to the address. With these data, the repository of places can provide all required information for the indexing stage (4), in which a spatial index (geoindex) is built to allow access to pages that include some geographic context. This kind of index represents an improvement over traditional indexes, used by search engines, since it uses types of geospatial evidence to determine a connection to a city, even when city and state names are left implicit in the text of the Web pages. For more details on spatial indexes, which are outside scope of this paper, see [17].

4.2 Patterns for Address Extraction

We defined address patterns based on three assumptions. First, every service on the Web includes a way for it to be located, so, even if an address is not provided, at least a contact phone number is present. Second, service pages in sites with greater credibility and better contents are usually more elaborate, so a well-formatted address is expected. Third, the identification of a postal code is an indication of the presence of a postal address.

In order to avoid false positives and ambiguities, four strategies have been adopted in the recognition of an address: (1) a phone number is only considered if it is accompanied by the respective area code; (2) a postal code that does not follow the defined pattern is only considered if the numbers are preceded by an adequate keyword (example: “ZIP:”, or similar); (3) a city name must always be followed by a state name or standard abbreviation; (4) every address must be accompanied by one or more location indicators (phone number, city/state, or postal code). The first strategy intends to avoid ambiguities among phone numbers, the second avoids false positives between a postal code and any other number, the third avoids ambiguities between city names, and the fourth avoids postal addresses with unknown location. Using these strategies, the identification of city and state is ensured.

After an exhaustive visual assessment of Brazilian Web pages containing addresses, we detected a large number of variations. Thus, it was necessary to define basic patterns, corresponding to the most important address parts: (1) *Basic Address*, (2) *Postal Code*, (3) *Phone Number*, and (4) *City/State*. The complement was not included as a basic pattern since it is optional, and varies considerably both in size and in contents.

We combined these four basic patterns in every possible way, and discarded unusual and counterintuitive combinations. The remaining 18 patterns were translated into regular expressions and implemented using PERL. Notice that the regular expressions can be automatically obtained from the contents of the ontology. From the regular expressions, a generic extractor should be capable of extracting addresses from any Web pages. A preliminary experiment was devised to identify the most efficient of these extraction patterns. In this experiment, each of the 18 patterns was tested over a collection of 75,413 pages from the Brazilian Web. This collection has been formed by automatically submitting to Google a representative set of landmarks and some generic references, such as “airport”, “downtown” and “bus station”. All pages that returned were collected and preprocessed. In this preliminary experiment, two questions were raised: (1) What is the most common format for addresses in Brazilian Web pages? (2) Which of the location identifiers appears more frequently? Our goal was to reduce the number of patterns to be submitted to a larger collection, to avoid redundant or ineffective patterns, keeping only the most representative ones. After the extraction, 893,260 addresses were recognized in 43,121 of the 75,413 pages, that is, 57% of the pages contained one or more addresses formatted according to one of the 18 patterns. The high percentage of pages with recognized addresses reflects the characteristics of the collection, thus confirming and validating its usefulness for the experiment.

Given the results from this experiment, the original 18 patterns have been reduced to 11, corresponding to the most successful patterns in address extraction (99.6% of all occurrences within the 75,413 pages) (Table 1). During a validation effort, carried out using a method similar to the one we present in Section 4.3, we observed that a few shorter patterns were actually contained on some of the larger ones, and thus retrieved the same addresses. We could, therefore, eliminate patterns 3 and 6, since they are redundant with pattern 5, and because geocoding using the postal code was more efficient than using phone number or city/state. Patterns 7 and 8 are redundant with patterns 9 and 10, and were likewise eliminated. Even though the use of city/state led to many false positives, we opted to keep pattern 4, considering the possibility of finding addresses with neither a postal code nor a phone number. Pattern 2 was kept as well, since we noticed the use of the postal code combined with city/state would make geocoding more reliable. Finally, pattern 11 was eliminated because of the large number of false positives, mostly caused by state abbreviations near expressions that are not city names. The large rate of false positives reduced the success in geocoding (50.76%), thus negating the apparently positive effect of a high extraction rate. We observed that the presence of one of the numeric location identifiers (postal code or phone number) ensures more efficient geocoding, since they are not as prone to spelling mistakes and abbreviations as the city/state identifier. As a result, patterns 1, 2, 4, 5, 9, and 10 were selected to be used in the next experiment.

Table 1 – The Most Effective Address Recognition Patterns

Patterns		Extractions	
1	*Basic Address + Phone Number*	26,350	2.9%
2	*Basic Address + City/state + Postal Code*	6,981	0.8%
3	*Basic Address + Postal Code + City/state*	21,424	2.4%
4	*Basic Address + City/state*	19,067	2.1%
5	*Basic Address + Postal Code*	43,074	4.8%
6	*Basic Address + Postal Code + Phone Number*	1,860	0.2%
7	*Phone Number + City/state*	858	0.1%
8	*Postal Code + Phone Number*	2,358	0.3%
9	*Phone Number*	151,641	17.0%
10	*Postal Code*	99,842	11.2%
11	*City/state*	516,413	57.8%
Subtotal		889,868	99.6%
Recognized addresses total		893,260	100.0%

4.3 Evaluation of the Extraction Patterns

The evaluation of the extraction patterns was divided into two steps: (1) contents assessment (Was the extraction correct?) and (2) extraction capability assessment (How much of the available information was really extracted?). These two steps allow the evaluation of the effectiveness of the extraction, and correspond, respectively, to two metrics widely used in information retrieval: precision and recall [2].

Contents assessment was performed in three steps. In the first step we geocoded the location identifiers, checking our Locus gazetteer [16] for the existence of each one. If a match was found, there was a big chance that the extraction was performed correctly. The second step selected, within a sample, all the extractions that were not geocoded in the first step, manually checking the extraction contents to determine whether the problem occurred at the extraction or at the geocoding. The third step verified the quality of the extracted addresses. This was done by comparing two sets of geocoding results: one using both the basic address and the city/state, and other using only city/state, as in the first strategy. If the geocoding of the basic address plus city/state is successful, it will return a point within the expected city's limits.

In the contents assessment, 385 extractions were randomly selected for each pattern (Table 2). All results were verified visually. A manual inspection of the geocoding results from the first step verified that, for the patterns including a phone number, failures resulted from outdated numbers, with a wrong area code or presented in an unusual format. In patterns including city/state most problems were in the recognition of the city name. The main issues include (1) lack of a separator between neighborhood name and city name; (2) cities that have the same name as the state they are in; (3) landmark names used where a city name was expected; (4) abbreviations in the city name. In the case of the postal code, a problem was caused by one of the regular expressions, which allowed the recognition of codes in a form that led to confusion with other numeric formats. As a result, numbers that are not related to postal codes were retrieved, such as IP addresses, population counts, and others. To avoid this problem, a part of the regular expression was eliminated from the extraction patterns.

Table 2 – Geocoding Results from Location Identifiers

Patterns	Extractions	Not Geocoded Extractions	
		Total	Correctly Recognized
Phone Number	385	123	120
Basic Address + City/State + Postal Code	385	0	-
Basic Address + Phone Number	385	116	115
Basic Address + City/State	385	192	3
Basic Address + Postal Code	385	5	5
Postal Code	385	26	4
Total	2310	462	247

Once the location identifiers were extracted, it was necessary to verify whether the basic addresses were extracted correctly. For this, we geocoded all addresses from eight distinct Brazilian cities that were found in the experiment. Results were quite satisfactory, and showed the patterns used for address extraction work well. Only a few addresses could not be geocoded precisely. From the 1092 addresses obtained from the sample, 94.6% were exactly geocoded (see Section 4.1), 1.8% were geocoded approximately, and 2.5% were not found. The remaining 1.1% could not be geocoded because the corresponding addresses could not be found in the geographic database we used, but all street names were found in the Brazilian Postal Services on-line street catalog.

After the patterns were verified, we moved on to the extraction capability assessment, in which all 385 pages of the sample were manually inspected, visually identifying all addresses that should have been recognized by the extractor. For each page, we generated a list of the patterns found and their frequencies. These results were compared to the actual extraction, thus calculating the percentage of addresses that have been automatically recognized, and indicating the causes for the failures.

Results show the percentage of addresses found with each of the six selected patterns was quite satisfactory. The best results correspond to the *postal code* (99.18%), followed by *Basic Address + Postal Code* (87.13%) and by *Basic Address + City/State* (73.91%). *Phone Number* obtained a slightly lower percentage (73.35%), considering the occurrence of fax numbers and several phone numbers associated to a single area code indication in some pages. This happened because the extractor would recognize only the first number found. The patterns *Basic Address + Phone Number* and the *Basic Address + City/State + Postal Code* extracted respectively 68.48% and 58.40% of the addresses. These experiments showed the feasibility of the idea of using addresses found on Web pages as reliable location indicators.

5. EXPERIMENTAL EVALUATION

The extractor with the six selected patterns was then applied to the WBR05 collection, which is composed by over 4 million pages, collected from Brazilian Web sites in March 2005. This collection significantly reflects the Brazilian Web.

As a result, 2,137,601 addresses were located on 603,798 pages (14.77% of the collection), confirming previous findings [9, 13]. This result shows that the Web is a rich source for local content, and that urban addresses, used as access keys to such content, are important assets, especially when the content refers to daily activities and services of local interest. In WBR05, 12% of the pages include one or more phone numbers, 6.99% include postal codes, and 9.53% include addresses (Table 3). Postal codes were the most effective geocoding resource, confirming the results from the preliminary experiment. There are about twice as many phone numbers in the collection as there are postal codes, reinforcing the notion that many pages only present phone and fax numbers as contact information. The pattern *Basic Address + City/State* showed the lowest geocoding percentage (77.07%), a little under the results for *Basic Address + Phone Number* (79.89%), for the same reasons discussed earlier. For the pattern *Basic Address + City/State + Postal Code*, the success rate in geocoding using the postal code only was much higher (99.33%) than when using only city/state (68.20%). This result validates our decision to keep this pattern among the six selected from the preliminary experiment. Due mostly to the already discussed imprecision in the extraction of city/state, in 33.25% of the cases there was a discrepancy between the geocoding from the postal code and from the city/state. In those cases, the results from using the postal code were more reliable. Table 4 shows, for each of the six patterns, the number of extracted and geocoded addresses.

Table 3 – Number of Pages that Include Each Pattern

Patterns	Number of Pages	
Phone Number	505,189	12.0%
Basic Address + City/State + Postal Code	24,475	0.60%
Basic Address + Phone Number	55,244	1.35%
Basic Address + City/State	5,063	3.79%
Basic Address + Postal Code	154,761	3.79%
Postal Code	285,999	6.99%

Table 4 – Number of Extracted and Geocoded Addresses

Patterns	Extracted	Geocoded	City/State Geocoded
Phone Number	1,083,913	79.89%	
Basic Address + City/State + Postal Code	34,832	99.33%	68.20%
Basic Address + Phone Number	99,297	81.46%	
Basic Address + City/State	217,274		77.07%
Basic Address + Postal Code	231,406	99.33%	
Postal Code	470,879	96.28%	

This experiment showed the feasibility of automatically recognizing, extracting, and geocoding addresses from Web pages. The results were once again satisfactory, showing that the six patterns selected from the first experiment are sufficient for address recognition. Using patterns eliminates the need to examine the textual vicinity of each known term to determine whether it is part of an address or not. Combining location identifiers with the basic address improved the precision of the result, reducing the number of false addresses. It was also possible to obtain a snapshot of the incidence of address information in Brazilian Web pages.

6. CONCLUSIONS AND FUTURE WORK

One of the main goals of geospatial evidence recognition is, among other applications, allowing the creation of mechanisms to enable search engines to perform local and proximity searches, without having to resort to yellow page directories. Our experiments have shown the feasibility of performing automated extraction and geocoding of addresses to identify locations associated to Web pages. Combining location identifiers with basic addresses improved the precision of the extractions, reducing the number of false positive results.

This paper focused on the local Web and presented an approach based on an ontology of urban places that allows recognition, extraction, and geocoding of geospatial evidence with local characteristics. We described experiments that evaluate the presence and incidence of urban addresses in Web pages. From the experiments, we showed that addresses provide satisfactory support for local search applications, since they represent the physical location of services and activities found in Web pages.

Our approach to achieve such goals is based on the recognition of addresses found within Web pages, and considers two levels of granularity: city (general identification of the city) and local (precise location of the address within the city). With city granularity, as presented in this paper, at least the association of a page of interest to a city is ensured, and geocoding coordinates resulting from this association can be used for proximity searches. A method to identify the most common patterns for address extraction was presented, and a minimal set of patterns for the extraction of Brazilian addresses was obtained and validated experimentally using a collection of over 4 million Web pages.

Results of this approach open perspectives for new types of useful applications which simplify, improve, and enhance local Web searches. Future work involves identifying service pages, categorizing services associated to extracted addresses, and associating the name of the service provider to the address.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

1. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-a-Where: Geotagging Web Content. In *Proc. of the 27th Annual Int'l ACM SIGIR Conf. on Res. and Develop. in Information Retrieval*. Sheffield, UK, 2004, pp. 273-280.
2. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley: Reading, MA, 1999.
3. Borges, K.A.V. Use of an Ontology of Urban Places for Recognition and Extraction of Geospatial Evidences on the Web (in Portuguese). PhD Thesis, Federal University of Minas Gerais: Belo Horizonte (MG), Brazil, 2006.
4. Borges, K.A.V., Laender, A.H.F., Medeiros, C.B., Silva, A.S., and Davis Jr., C.A. The Web as a Data Source for Spatial Databases. In *Proc. of the V Brazilian Symp. on GeoInformatics*. Campos do Jordão (SP), Brazil, 2003, pp. CD-ROM.
5. Davis Jr., C.A. and Fonseca, F.T. Assessing the Certainty of Locations Produced by an Address Geocoding System. *Geoinformatica*, **11**(1): 103-129, 2007.
6. Delboni, T.M., Borges, K.A.V., Laender, A.H.F., and Davis Jr., C.A. Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. *Transactions in GIS*, **11**(3): 377-397, 2007.
7. Embley, D.W., Campbell, D.M., Jiang, Y.S., Liddle, S.W., Lonsdale, D.W., Ng, Y.-K., and Smith, R.D. Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. *Data & Knowl. Eng.*, **31**(3): 227-251, 1999.
8. Fu, G., Jones, C.B., and Abdelmoty, A. Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proc. of the IASTED Int'l Conf. on Databases and Applications*. Innsbruck, Austria, 2005, pp. 167-172.
9. Himmelstein, H. Local Search: The Internet is the Yellow Pages. *IEEE Computer*, **38**(2): 26-35, 2005.
10. Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Van Kreveld, M., and Weibel, R. Spatial Information Retrieval and Geographical Ontologies: an overview of the SPIRIT project. In *Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Res. and Develop. on Information Retrieval*. Tampere, Finland, 2002, pp. 387-388.
11. Laender, A.H.F., Ribeiro-Neto, B.A., Silva, A.S., and Teixeira, J.S. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record*, **31**(2): 84-93, 2002.
12. Larson, R.R. Geographic Information Retrieval and Spatial Browsing. In *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, Smith, L.C. and Gluck, M., Eds. 1996, Un. of Illinois: Urbana, IL. p. 81-123.
13. McCurley, K.S. Geospatial Mapping and Navigation on the Web. In *Proc. of the Tenth Int'l World Wide Web Conference (WWW10)*. Hong Kong: ACM, 2001, pp. 221-229.
14. Sanderson, M. and Kohler, J. Analyzing Geographic Queries. In *Proc. of the ACM SIGIR Workshop on Geographic Information Retrieval*. Sheffield, UK, 2004, pp. 1-2.
15. Silva, M.J., Martins, B., Chaves, M., Cardoso, N., and Afonso, A.P. Adding Geographic Scopes to Web Resources. *Computers, Environment and Urban Syst.*, **30**: 378-399, 2006.
16. Souza, L.A., Davis Jr., C.A., Borges, K.A.V., Delboni, T.M., and Laender, A.H.F. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *Proc. of the 3rd Latin American Web Congress*. Buenos Aires, Argentina, 2005, pp. 157-165.
17. Vaid, S., Jones, C.B., Joho, H., and Sanderson, M. Spatio-textual Indexing for Geographical Search on the Web. In *Proc. of the 9th Int'l Symp. on Spatial and Temporal Databases*. Angra dos Reis, Brazil, 2005, pp. 218-235.
18. Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W. Detecting Geographic Locations from Web Resources. In *Proc. of the 2nd Int'l Workshop on Geographic Information Retrieval*. Bremen, Germany, 2005, pp. 17-24.
19. Zong, W., Wu, D., Sun, A., Lim, E., and Goh, D.H.G. On Assigning Place Names to Geographic Related Web Pages. In *Proc. of the 5th ACM/IEEE-CS Joint Conf. on Digital Libraries*. Denver, Colorado, USA, 2005, pp. 354-362.