International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 5 (2013) pp. 089-097
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Geographic Information Retrieval and Visualization of Online Unstructured Documents

Rocío Abascal-Mena[1], Erick López-Ornelas[1] and J. Sergio Zepeda[1]

[1] Universidad Autónoma Metropolitana – Unidad Cuajimalpa,
Departamento de Tecnologías de la Información, Av. Constituyentes 1054, 5° piso, Col. Lomas Altas
Del. Miguel Hidalgo, C.P. 11950, México D.F., México
mabascal@correo.cua.uam.mx, elopez@correo.cua.uam.mx, jzepeda@correo.cua.uam.mx

***Abstract*: Newspapers, travel narratives, blogs, books and the Internet hold a huge amount of geographic information that can be extracted in order to provide visual exploration. Also, the understanding of place references involves knowledge of the document context. In this way, the study of tools for disambiguation is needed. For the automatic annotation of time and location, both shared world knowledge and document context needs to be captured. This paper is centered on analyzing online unstructured documents: travel narratives and online newspapers. Our approach is based on the exploration of tools able to make automatically the disambiguation of placenames. In this case, we have used a Geoparsing Web Service to extract geographic coordinates from the online unstructured documents. Once geographic coordinates are extracted by using eXtensible Markup Language (XML) we draw the geo-positions and link documents into a map image in order to visualize textual information.**

Keywords: information extraction, visualization, semantic web, Geoparsing Web Services, georeferencing, online newspapers, online travel stories.

I. Introduction

Due to the popular use of the Internet and the dramatic progress of telecommunication technology, the paradigm of Geographic Information Systems is shifting into a new direction, especially for producing spatial-oriented visualization [1]. In this context, the ability to process or search data and enrich it using other visual resources, such as maps, in acceptable time, becomes a powerful asset. In our days, we find more and more use of the Web in order to share information with social groups. But, there is a lack on the capacity to find and extract this type of information. Furthermore, one of the characteristics of information provided by *"common users"* is geo-referenced information, meaning that we have an absolute position, such as the pair longitude-latitude, introducing location-aware concepts to information retrieval systems. Also, a great amount of people is using new media (cell phones, PDAs and smart phones) to consult news online.

Nowadays, navigating the Web has become a common way to find information needed to solve everyday problems [2]. So, when people browse the web for information, they must base navigation decisions on what they read and of course visual information can help in taking these decisions. That's why we are interested in providing new services to enable more rapid consulting of the existing information. Our study starts in the analysis of resources already found in the Web: travel narratives and newspaper's covers.

In the case of travel narratives, we found that they are a type of geo-referenced information which introduces a new opportunity to improve location-aware services. If we consider that this information has a time stamp associated, as the newspapers, a place results represented by a list of concepts that changes in time.

In the case of online newspapers covers, we extract the places coming from them in order to determine its importance during a scale of time and allow the user to extract the title of news corresponding to a concept selected. The newspaper covers, as the travel narratives, provide geographical information that can be elicited in order to analyze the places that are most covered over the World. To make this possible, we have explored the different web mapping services existing in this specific application.

The article is structured as follows. Some background of Geoparsing Web Services (GWS) and geo-semantics is provided in Section 2. Section 3 describes the methodology followed to identify visual elements in online unstructured documents. In Section 4 we describe the first case of analysis: travel narratives. Next, in Section 5, we describe a second case of analysis concerning online newspapers. Experimental results are also described. Some conclusions and further work are shown in Section 6.

II. Background

Web mapping services applications like Google Maps, Google Earth, NASA World Wind or Flickr map are a way to organize the world's information geographically. Mapping services have been used in many areas including weather forecast, tourism and asset management. They provide geospatial visualization of information so the users can analyze, plan and take decisions based on geographic location. These services help users to understand the relationship between data and geographic location. All mapping applications provide an intuitive mapping interface with detailed street and aerial imagery data embedded. In addition, map controls can be embedded in the product to give users full control over map navigation.

The primary goal behind its rapid acceptance as an Internet mapping viewer is the ability to customize the map to fit application specific needs. Although, there is a lack in the use of these services in order to apply them in the extraction of geographic concepts coming from unstructured texts. In this way, we have to use *geoparsing* which is the process of recognizing geographic context [3]. The first step involves extracting geographic entities from texts and distinguishing them from other entities such as names of people or organizations, and events. In natural language processing this is referred to as Named Entity Recognition (NER) and is central to other text processing applications such as information extraction (IE) and information retrieval (IR).

Geoparsing is most frequently used to automatically analyze collections of text documents. There are a number of commercial products with a geoparsing capability. Companies like MetaCarta extract information about place and time, while others like Digital Reasoning (GeoLocator), Lockheed Martin (AeroText), and SRA (NetOwl) extract places along with other entities, such as persons, organizations, time, money, etc. To process the large volumes of data, these systems rely on automated techniques optimized for speed. These geoparsing systems are not perfect. Identifying and disambiguating place names in text are difficult and vulnerable to the vagaries of language. Just identifying which words are associated with place names can be a challenge. The geoparsing software must not only understand the words, but whether the words that form a name actually refer to a place. The software must understand that "Paris" in "Paris, France" refers to an urban area; in "Paris Creek" refers to a stream; in "Paris Hilton" refers to a person or to a hotel; and in "Paris Match" refers to a magazine name. But, once a place name has been identified, disambiguation remains a difficult. For example, there are over 2,100 names in the National Geospatial-Intelligence Agency which exactly match San Antonio. Sometimes, without being the author of a document, it is not possible to identify, with any confidence, the place to which a name refers.

Geographic Information Retrieval (GIR) has also emerged as an active and growing research area, addressing the retrieval of collections of textual documents according to geographic criteria of relevance [4]. But, rather than focus on analyzing collections of documents, some other approaches focuses on the individual document, allowing authors to efficiently ensure that the place names are identified correctly and are discoverable by other users. Just as map documents go through a review and validation process, this approach allows authors to confirm that the places in their documents are correctly identified and located at the time of writing. One example is GeoDoc where the user has to identify and tag the place names manually, the application starts by automatically extracting place names and highlighting them on the display.

Current work on query processing for retrieving geographic information on the Web sometimes requires a combination of text and spatial data techniques for usage in geographic web search engines. A query to such an engine consists of keywords and the geographic area the user is interested in (i.e., query footprint).

We find other works referring to the use of geographic locations in order to link pages to which they are related [5]. In this work, they also assign to each page a geographic focus that is retrieved by the content the page discusses as a whole. Furthermore, their *"tag enrichment"* process consists of finding place entities that show potential for geo-referencing, and then applying a disambiguation taxonomy.

A. The Yahoo! Placemaker Web Service

Geographic text mining technology is nowadays mature and commercial services offering geoparsing functionalities, for the identification of places, are starting to appear [4]. That is the case of the Yahoo! Placemaker which is a geotagging web service that provides third-party developers the means to enrich their applications or Web sites with geographic information. The service is able to identify, disambiguate, and extract place names from unstructured and semi-structured documents. It is also capable of using the place references in a document, together with a pre-determined set of rules, to discover the geographic scope that best encompasses its contents. Thus, given a textual document, Yahoo! Placemaker returns unique *Where-on-Earth Identifiers* (WOEIDs) for each of the named places and scopes. Using these identifiers it is possible to query the Yahoo! GeoPlanet gazetteer service, and obtain further information on the location. This way, each of the resolved place references is associated with the corresponding city, state, country and continent, as well as with the bounding rectangle that covers its area.

There are two flavors of document scopes in Placemaker, namely the geographic scope and the administrative scope. The geographic scope is the place that best describes the document. The administrative scope is also the place that best describes the document, but is of an administrative type (i.e., Continent, Country, State, County, Local Administrative Area, Town, or Suburb). Since the reference document collection that we used for our experiments only contains documents assigned to administrative regions, we limited our cross-method comparison to using Placemaker's administrative scopes.

The Yahoo! Placemaker Web Service is a commercial product and not many details are available regarding its functioning. However, some information about the service is available in the Web site, together with its documentation. For instance, the Web site claims that when the service encounters a structured address, it will not perform street level geocoding but will instead disambiguate the reference to the smallest bounding named place known, frequently a postal code or neighborhood. The Web site also claims that besides place names, the service also understands geography-rich tags, such as the W3C Basic Geo Vocabulary and HTML microformats. However, no details about the rules that are used in the scope assignment process are given in the documentation for the service.

The Placemaker Web service accepts plain text as input, returning an eXtensible Markup Language (XML) document with the results. The service has an input parameter that allows users to provide the title of the document separately from the rest of the textual contents, weighting the title text as more representative. In our experiments, we used the Web service as a black-box to assign scopes to the Web documents, using the option that weights the title text as more important than the rest.

However, we find some work that has been done by using the Placemaker, this is the case of Tobin et al., [6] which makes a comparison of their system with the Placemaker. They notice a high-precision recognition system, in the Placemaker, which might well out-perform their work. However, for some places, the Placemaker has a low coverage issue like is the case of the Chinese location [7]. Some experiments, using the Placemaker, have been done in order to extract place references from the text and assign each document to a corresponding geographic scope [4], [8].

We present a methodology for extracting semantic information about places, from two cases: 1) travel narratives and 2) online newspapers. By extracting semantic knowledge from places, it becomes possible to have a view on the dynamic life of travel narratives and newspapers covering cities, countries, etc.

III. Methodology

Our approach is based on the use of a Geoparsing Web Service (GWS) which enriches content with geographic metadata by extracting places from unstructured texts, online travel narratives and online newspaper covers. Geoparsing offers the ability to turn text documents into geospatial databases. This process is done in two steps: 1) entity extraction and 2) disambiguation, which is also known as grounding or geotagging. Geospatial entity extraction uses natural language processing to identify place names in text, while disambiguation associates a name with its correct location.

In order to access the GWS we have used the Yahoo! Placemaker, which is a GWS that provides third-party developers the means to geo-enrich content at scale. The service identifies, disambiguates, and extracts places from unstructured and structured textual content: web pages, RSS (and Atom) feeds, news articles, blog posts, etc. It is an open API that assists developers in creating local and location-aware applications and datasets. Placemaker is a geo-enrichment service that assists developers in determining the whereness of unstructured and atomic content, making the Internet more location-aware.

To access the GWS we have used the Yahoo Query Language (YQL) which is a query influenced by the Structured Query Language (SQL) but diverges from it as it provides specialized methods to query, filter, and join data across web services. The process is shown in Figure 1.

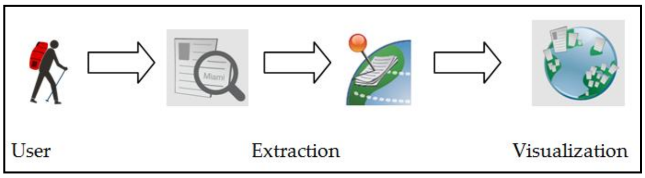

Figure 1. Geoparsing and visualization process.

In the next section, we present the first case of analysis: travel narratives or stories coming from the Web.

IV. The Case of Travel Narratives Analysis

On the web we find many websites containing travel stories. But, to present our experimental results we have chosen only one of these sites. We selected "The Adventure Prone Site" that can be found in: http://www.adventureprone.com/. From this site, we have extracted all the URL containing travel narratives by using the web spider called Robot V1 (http://www.semantic-knowledge.com/). This spider is designed to collect websites and extract texts from Internet, following the links from a starting Web page to other pages, until the process is finished. Thus, we have obtained pages coming from this site. To analyze and extract the geographic aspects of these pages we have implemented a system able to communicate to the GWS by using YQL. For each URL extracted we have obtained the geographic coordinates. But, in order to present our results, we decided to choose only one travel story: "A Tale of Ten Cities" which is found here: http://www.adventureprone.com/travel/stories/cities.html. Analyzing this page, we have obtained 48 places with the next geographic elements: name (latitude, longitude). These places are shown below.

1. Budapest, Budapest, HU (47.5062, 19.0648)
2. San Francisco, CA, US (37.7792, -122.42)
3. France (46.7107, 1.71819)
4. Silicon Valley, CA, US (37.3953, -122.053)
5. Italian Town, AL, US (33.1183, -87.1)
6. Danube, HU (46.3298, 18.9073)
7. Poland (51.9189, 19.1343)
8. Naples, Campania, IT (40.8399, 14.2519)
9. Gary, Midi-Pyrénées, FR (43.6959, 1.91942)
10. Greece (39.0724, 21.8456)
11. Prague, Hlavni mesto Praha, CZ (50.0791, 14.4332)
12. Armenia (40.0662, 45.0399)
13. Acropolis, Athens, Attiki, GR (37.9714, 23.7238)
14. England, GB (52.8836, -1.97406)
15. Mont Blanc, Bossons, Rhône-Alpes, FR (45.8359, 6.86211)
16. Berlin, Bundesland Berlin, DE (52.5161, 13.377)
17. Pantheon, Rome, Lazio, IT (41.8987, 12.4769)
18. Colosseum, Rome, Lazio, IT (41.8902, 12.4929)
19. Montpellier, Languedoc-Roussillon, FR (43.6109, 3.87609)
20. Rome, Lazio, IT (41.9031, 12.4958)

21. Italy (42.5038, 12.5735)
22. Spain (39.895, -2.98868)
23. Venice, Veneto, IT (45.4383, 12.3185)
24. Bolivia (-16.2883, -63.5494)
25. Manarola, Liguria, IT (44.1075, 9.73006)
26. Chamonix-Mont-Blanc, Rhône-Alpes, FR (45.9249, 6.87193)
27. Florence, Tuscany, IT (43.7824, 11.255)
28. Pisa, Toscana, IT (43.71, 10.3995)
29. London, England, GB (51.5063, -0.12714)
30. Vesuvius, Torre del Greco, Campania, IT (40.8, 14.4)
31. Newquay, England, GB (50.4158, -5.07558)
32. Monastiraki, Athens, Attiki, GR (37.9782, 23.7268)
33. St Peter's Basilica, Vatican City, VA (41.9022, 12.4547)
34. Vatican Museums, Vatican City, VA (41.9069, 12.454)
35. St Peter's Square, Vatican City, VA (41.9023, 12.4576)
36. Athens, Attiki, GR (37.9762, 23.7364)
37. United Kingdom (54.3141, -2.23001)
38. St. Columb Major, England, GB (50.4325, -4.93688)
39. Vatican City (41.9038, 12.4525)
40. Morocco (31.8154, -7.067)
41. Buda, Budapest, Budapest, HU (47.5131, 19.0241)
42. Africa (2.07079, 15.8005)
43. Trevi Fountain, Rome, Lazio, IT (41.9009, 12.4833)
44. Cairo, Al Qahirah, EG (30.0837, 31.2554)
45. Olympia, Olimbia, Dytiki Ellada, GR (37.6405, 21.6281)
46. Europe (52.9762, 7.85784)
47. London Stansted International Airport, Takeley, England, GB 51.8894, 0.26256)
48. Delphi, Dhelfoi, Sterea Ellada, GR (38.472, 22.4746)

In the previous results presented we find places like *"San Francisco, CA, US"* which is compared in the story even if it is not part of the trip: *"We would go out to dinner every night, and seeing as things were so "cheap" compared to San Francisco, spend US$50/head on a fantastic meal with great wines which would have cost twice as much at home."* This is an example of further work that must be done in order to extract only the visited places according to the context.

With the geographic coordinates obtained we display each position on an Equirectangular projection of the Earth. For example, for the 48 coordinates extracted we have obtained the projection show in Figure 2.

To present in a clearly way the results obtained, we have selected only the coordinates belonging to Europe to show them in an Equirectangular Europe projection. In this way, Figure 3 presents the places visited in Europe during the tour written in the analyzed travel story.

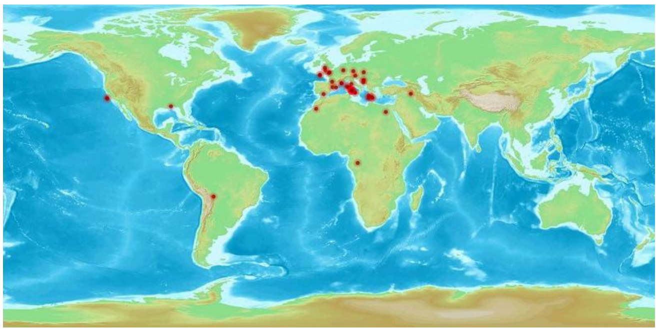

Figure 2. Representation of 48 coordinates extracted from the travel narrative: "A Tale of Ten Cities"

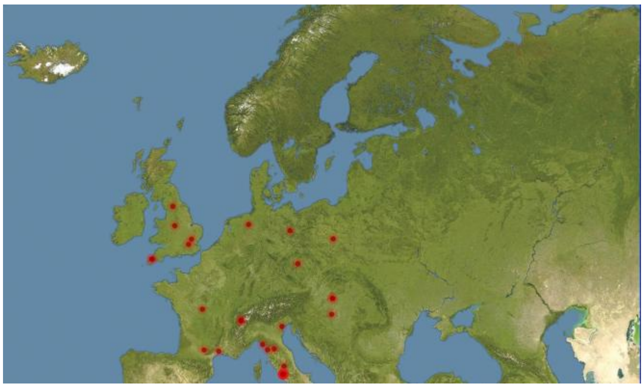

Figure 3. Representation of extracted cities corresponding to Europe.

Each of the places linked to the map contains its respective paragraph where the place it's mentioned. Furthermore interaction possibilities from Geographic Information Systems like zooming, filtering, buffering could be implemented, even to represent this map as a layer on a regular GIS.

In the next section, we present the second case of analysis: online newspapers covers.

V. The Case of Online Newspapers Analysis

The central thesis of our work is that geospatial information provides an important source of structure that can be directly integrated with textual content for organizing global news.

As researchers discovered a decade ago with large-scale collections of Web pages studying the connective structure of a corpus at a global level exposes a fascinating picture of what the world is paying attention to. In the case of global news, it means that we can discover, through newspapers covers, what people consider to be the most significant news in the world and within specific countries or cities, which of them figured the most in the attention of journalists all over the world. These resulting views of the data add to an emerging theme in which planetary-scale datasets provide insight into different kinds of human activity.

A. Dataset and Experimental Results

Our dataset was collected by extracting, automatically, the name of the places coming from the main covers of the world's newspapers: *New York Times (USA), Le Monde (France), Morgen (Belgium), ABC (Spain), Il Giornale (Italy), Haberturk (Turkey), Tages Anzeiger (Swiss), Dagens Hyheter*

(Sweden), Telegraph (United Kingdom), Jornal de Noticias (Portugal), AD (Netherlands), Irish Examiner (Ireland), Gulf News (Arab Emirates), The Times of India (India), Haaretx (Israel), Sydney Morning Herald (Australia), Asahi (Japan), Reforma (Mexico), La Razón (Argentina), El Comercio (Peru), The Times (South African) and *The Mercury (South African).*

Our goal was to retrieve as large and unbiased a sample of georeferenced places as possible. To do this, for each newspaper cover we have extracted, by using the Geoparsing Web Service, the exactly name of the place, its latitude and its longitude. For example for the next news *"Water Pumping Begins at Japan Nuclear Reactor"* taken from the New York Times (http://www.nytimes.com/, Tuesday, April 19, 2011) the extraction obtained is: **Japan (37.4876, 139.838)**. In this case, 37.4876 correspond to the latitude and 139.838 to the longitude.

In order to show the importance of our research, we present, an example of the results of a process that has been done. In this case, we have extracted all the places covered in the newspapers during a period of 3 weeks. We have selected one newspaper, the Morgen (Belgium), to show the places extracted during 3 different days of the 3 weeks analyzed: March 30, (Table 1), April 6 (Table 2) and April 13, 2011 (Table 3).

Morgen	**March 30, 2011**
Belgium	Anderlecht, Brussels, Capital Region of Brussels, BE (50.829, 4.29247)
	Bahrain (26.0247, 50.5485)
	Ban Doet, Yasothon, TH (15.9539, 104.063)
	Banyoles, Catalonia, ES (42.117, 2.76498)
	Berlin, Berlin, DE (52.5161, 13.377)
	Catalonia, ES (41.6922, 1.74161)
	Deynze, Oost-Vlaanderen, BE (50.987, 3.52924)
	Fukushima Prefecture, JP (37.384, 140.104)
	Germany (51.1642, 10.4542)
	Hasselt, Limburg, BE (50.9271, 5.33598)
	Iran (32.4207, 53.6824)
	Japan (37.4876, 139.838)
	Libya (26.3385, 17.2688)
	Miami, FL, US (25.729, -80.2374)
	Netherlands (52.1082, 5.32986)
	Ottawa, Ontario, CA (45.4215, -75.6919)
	Ukkel, Brussels, Capital Region of Brussels, BE (50.7904, 4.3625)
	Zulte, Oost-Vlaanderen, BE (50.9216, 3.44363)

Table 1. Extracted places from Morgen (Belgium) on March 30, 2011.

The Table 1 show 18 places corresponding to the most important news consider in the Morgen newspaper. In this way, we can see the apparition of places like Japan **(37.4876, 139.838)** or **Libya (26.3385, 17.2688)** which they were having trouble in the moment of our analysis. News like: *"Japan Earthquake 2011: 8.9 Magnitude Earthquake Hits, 30-Foot Tsunami Triggered"* or *"In Libya, Muammar Gaddafi's forces have mounted attacks on both sides of the country, bombarding the city of Misurata in the west and pushing back the rebel advance in the east"* were the bread of today's news. Also, in others newspapers we find coincidences resulting from the interest of the people in these world's news.

In Table 2, we continue to find some places as in last week (Table 1) like Germany (51.1642, 10.4542) and Netherlands (52.1082, 5.32986).

Morgen	**April 6, 2011**
Belgium	Astana, Astana, KZ (51.1894, 71.4321)
	Basque Country, ES (42.9639, -2.58927)
	Blankenberghe, West-Vlaanderen, BE (51.311, 3.13432)
	Brussels, Capital Region of Brussels, BE (50.8484, 4.34968)
	Bucharest, Bucuresti, RO (44.4342, 26.103)
	Cavendish, VT, US (43.3828, -72.6068)
	Chicago, IL, US (41.8842, -87.6324)
	Cleveland, OH, US (41.5047, -81.6907)
	Denderleeuw, Oost-Vlaanderen, BE (50.8831, 4.0811)
	Dendermonde, Oost-Vlaanderen, BE (51.0315, 4.09794)
	Detroit, MI, US (42.3317, -83.0479)
	Freiburg, Baden-Wurttemberg, DE (47.9985, 7.84965)
	Germany (51.1642, 10.4542)
	Ghent, Oost-Vlaanderen, BE (51.0556, 3.72856)
	Indiana, US (39.7662, -86.441)
	Indianapolis Motor Speedway, Indianapolis, IN, US (39.7886, -86.2384)
	Indianapolis, IN, US (39.7669, -86.15)
	Interland, Davis, CA, US (38.5419, -121.728)
	Israel (31.3893, 35.3612)
	L'étape, Champagne-Ardenne, FR (48.351, 4.46286)
	Lithuania (55.174, 23.8944)
	Netherlands (52.1082, 5.32986)
	Russia (59.4538, 108.831)
	Ryckevorsel, Antwerp, BE (51.35, 4.75954)
	Saint-Trond, Limburg, BE (50.8153, 5.18637)
	Schoten, Antwerp, BE (51.251, 4.49799)
	Sudan (13.3166, 30.2095)
	Tamse, Oost-Vlaanderen, BE (51.1247, 4.2143)
	Tegen, Vastra Gotaland, SE (58.7618, 12.3471)
	Turkey (38.9577, 35.4317)
	Welkom, Free State, ZA (-27.9798, 26.7357)
	Zuya, Basque Country, ES (42.9562, -2.81859)

Table 2. Extracted places from Morgen (Belgium) on April 6, 2011.

Making a new comparison between the last two weeks (Table 2 and 3) we find coincidences in: Brussels, Capital Region of Brussels, BE (50.8484, 4.34968), Detroit, MI, US (42.3317, -83.0479) and Tegen, Vastra Gotaland, SE (58.7618, 12.3471). Also Libya reappears in the third week, as in the first one.

This way, we can recognize what newspapers want to cover more places and the ones that focus in only one percentage of the world's news.

In conclusion, the extraction of places coming from a set of newspapers covers, during a period of time, reflects the importance of world's news to different populations. For countries like USA, France, Belgium, Sweden, Netherlands, Japan and South African, the importance of covering the entire world's news is exemplified in the extraction that we have done. Instead, the newspapers from places like Israel, Turkey, Arab Emirates or even Peru are restricted to their own countries.

Morgen *Belgium*	**April 13**
	Abidjan, Lagunes, CI (5.32339, -4.02627) Antwerp, Antwerp, BE (51.2221, 4.39771) Berlin, Berlin, DE (52.5161, 13.377) ***Brussels, Capital Region of Brussels, BE (50.8484, 4.34968)*** Camembert, Lower Normandy, FR (48.8927, 0.17764) Detroit, MI, US (42.3317, -83.0479) India (21.7866, 82.7948) Ivory Coast (7.54685, -5.54709) ***Libya (26.3385, 17.2688)*** Long Island, NY, US (40.8525, -73.1358) Luxembourg (49.8152, 6.13348) Naples, Campania, IT (40.8399, 14.2519) Paris, Ile-de-France, FR (48.8569, 2.34121) Russia (59.4538, 108.831) Sharm el Sheikh, South Sinai, EG (27.8696, 34.3041) ***Tegen, Vastra Gotaland, SE (58.7618, 12.3471)*** Thielt, West-Vlaanderen, BE (51.0001, 3.32615)

*Table 3.*Extracted places from Morgen (Belgium) on April 13, 2011.

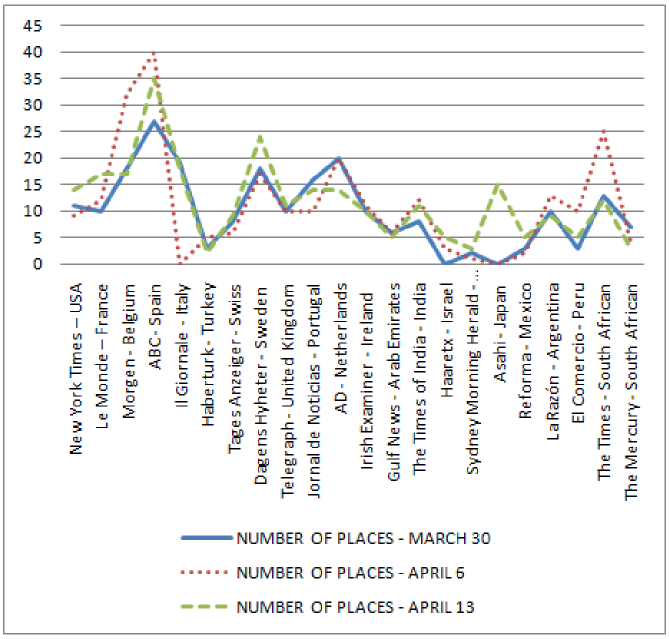

Figure 4. Number of extracted places from newspapers covers during 3 different weeks.

B. *Search of News based on Semantic Concepts*

As an additional example of our system we present the possibility to retrieve news based on the selection of semantic concepts. The user selects a concept and this is searched in all the newspapers covers. Our system presents the main news titles, for each newspaper, that corresponds to the concept searched, see Figure 5.

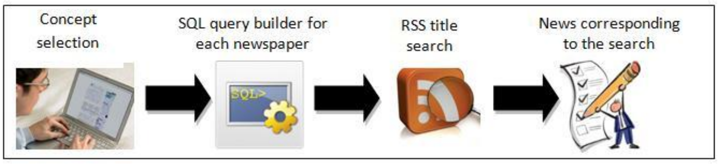
Concept selection
SQL query builder for each newspaper
SQL>
RSS title search
News corresponding to the search

Figure 5. Process to extract a list of news, each for every newspaper, containing the concept searched by the user.

Newspaper	**News title**
Le Monde *France*	**1:** L'audience qui a conduit DSK en prison **2:** Les images de DSK marquent "une terrible fin d'époque"
Morgen *Belgium*	**1:** Beelden geboeide topman IMF onmogelijk in België
ABC *Spain*	**1:** La UE y el FMI piden a la banca que colabore en el segundo rescate griego **2:** Strauss-Kahn será trasladado a la cárcel de Rikers Island
AD *Netherlands*	**1:** 'Geen immuniteit voor Strauss-Kahn' **2:** Strauss-Kahn zit vast tussen zware criminelen
The examiner *Ireland*	**1**: Kahn faces second assault claim **2:** Boyle under fire for Strauss-Kahn tweet
El Comercio *Peru*	**1:** Strauss-Kahn pasó de una suite de US$3 mil a una celda de 3x4 en 48 horas

Table 4. Newspapers with their corresponding titles containing *"Strauss-Kahn"* .

In order to represent the way the system works we have selected a concept, to be searched. In this case, at May 17, 2011 we found the next news *"The head of the International Monetary Fund, Dominique Strauss-Kahn, has been charged by New York police over an alleged sex attack on a hotel maid."* To know which of the newspapers talks about this news we decided to use the name *"Strauss-Kahn"* like a concept to make a search and automatically obtain all the titles, from all the newspapers at the same time, containing it. Table 4, shows the results of only the newspapers containing, in their cover, the concept *"Strauss-Kahn"*.

For the newspapers that appear in Table 4, we have extracted, also, all the places that they covered at the same time than the concept of *"Strauss-Kahn"* on May 17, 2011. In the Table 5 and the Table 6, we show the results.

In Table 5, we show the extracted places for the newspapers: Le Monde (France), Morgen (Belgium), El Comercio (Peru) and The Examiner (Ireland).

Newspaper	Extracted places for each newspaper
Le Monde *France*	New York, NY, US (40.7146, -74.0071) Beijing, Beijing, CN (39.906, 116.388) Marrakesh, الحوز ـ تـ انـ سـ يـ فت ـ مراكش, MA (31.6338, -8.00241) Ontario, CA (49.3843, -84.7563) Abidjan, Lagunes, CI (5.32339, -4.02627) Elío, Navarre, ES (42.7929, -1.80799) Ajaccio, Corsica, FR (41.9171, 8.73298) Syria (34.8148, 39.056) Moscow, Moscow Federal City, RU (55.7569, 37.6151) Rabat, زع ير-زمور- سلا-ال رب اط, MA (34.0209, -6.84165) Manhattan, New York, NY, US (40.7909, -73.9664) Ireland (53.4196, -8.24055) Pakistan (30.4419, 69.3597)
Morgen *Belgium*	Czech Republic (49.8039, 15.4749) Hainault, BE (50.3618, 3.89776) Berchem, North Brabant, NL (51.7743, 5.5656) United States (37.1679, -95.845) Cuba (21.511, -77.8068) Hong Kong (22.4112, 114.154) Moscow, Moscow Federal City, RU (55.7569, 37.6151) Pascal, Setif, DZ (35.8544, 5.291) Ghent, Oost-Vlaanderen, BE (51.0556, 3.72856) Kazakhstan (47.9992, 66.9023) Antwerp, Antwerp, BE (51.2221, 4.39771) Belgium (50.501, 4.47684) Brussels, Capital Region of Brussels, BE (50.8484, 4.34968) Toksook Bay Airport, Toksook Bay, AK, US (60.5413, -165.087)
El Comercio *Peru*	Lima, Lima Metropolitan Area, PE (-12.0436, -77.0212) Peru (-9.18134, -75.0024) Hasta, Maharashtra, IN (20.2837, 75.244) Ghana (7.95501, -1.03182)
The examiner *Ireland*	Portugal (39.5579, -7.84481) Dublin, Dublin, IE (53.3438, -6.24953) United Kingdom (54.3141, -2.23001) Northampton, England, GB (52.2397, -0.88576) Clare, IE (52.8639, -9.11365) Islamabad, Islamabad, PK (33.7098, 73.0759) Vatican City (41.9038, 12.4525) Ireland (53.4196, -8.24055) Washington, DC, US (38.8991, -77.029) Pakistan (30.4419, 69.3597) Muenster, North Rhine-Westphalia, DE (51.963, 7.61781)

Table 5. Extraction of places from the newspapers (Le Monde, Morgen , El Comercio, The Examiner) which have on May 17, 2011, news talking about *"Strauss-Kahn"*.

In Table 6, we show the extracted places for the newspapers: ABC (Spain) and AD (Netherlands).

Our system allows also the possibility to automatically visualize the places extracted for each of the newspapers analyzed. Figure 6 and 7, shows them for the case of the newspaper ABC (Spain).

Newspaper	Extracted places for each newspaper
ABC *Spain*	Puerto Lopez, Manabi, EC (-1.56239, -80.8058) Ireland (53.4196, -8.24055) Kerry, IE (52.1387, -9.51447) Palacio de Liria, Madrid, Madrid, ES (40.4267, -3.71341) Spain (39.8949, -2.98831) Chacón, Murcia, ES (37.7333, -1.0167) Rikers Island, New York, NY, US (40.7921, -73.8817) Pensiones, Toluca de Lerdo, Mexico, MX (19.2929, -99.646) Salida, CO, US (38.5366, -105.992) Malaga, Andalusia, ES (36.7183, -4.42016) Burgos, Castille and Leon, ES (42.3411, -3.69981) Libertad, Tachira, VE (7.8272, -72.3228) New York, NY, US (40.7146, -74.0071) Becerril, Castille and Leon, ES (42.1105, -4.64276) Madrid, Madrid, ES (40.4203, -3.70577) Dublin, Dublin, IE (53.3438, -6.24953) El Segundo, CA, US (33.9199, -118.416) Wentworth, England, GB (53.4789, -1.4186) Plataforma, Bahia, BR (-12.8943, -38.4841) Carme, Catalonia, ES (41.532, 1.62021) Reunion (-21.1145, 55.5321) Convivencia, Huimanguillo, Tabasco, MX (17.8312, -93.3826) Lorca, Murcia, ES (37.6805, -1.69056) Seville, Andalusia, ES (37.3877, -6.00181) Civica, Castille la Mancha, ES (40.8029, -2.7863) Ayer, Canton of Valais, CH (46.1775, 7.60374)
AD *Netherlands*	New York, NY, US (40.7146, -74.0071) Excelsior, San Francisco, CA, US (37.7211, -122.429) Moscow, Moscow Federal City, RU (55.7569, 37.6151) France (46.7107, 1.71819) Heerenveen, Friesland, NL (52.9587, 5.92882) Groningen, Groningen, NL (53.2171, 6.57356) Paris, Ile-de-France, FR (48.8569, 2.34121) Copenhagen, Hovedstaden, DK (55.6763, 12.5694) Rijswijk, North Brabant, NL (51.7977, 5.027) Rome, Lazio, IT (41.9031, 12.4958) Rotterdam, South Holland, NL (51.9228, 4.47845)

Table 6. Extraction of places from the newspapers (ABC and AD) which have on May 17, 2011, news talking about *"Strauss-Kahn"*.

VI. Conclusions and Further Work

In this paper, we presented a system for the extraction of geographic places coming from travel narratives and online newspapers. We enrich our system by allowing the search of stories and news according to semantic concepts. This capability can be useful to find only the story, the news or the newspaper in which the user is interested without having to read the entire document. Also, what is the biggest contribution of our work is that places covered in travel stories and newspapers can be visualized in an Earth projection.

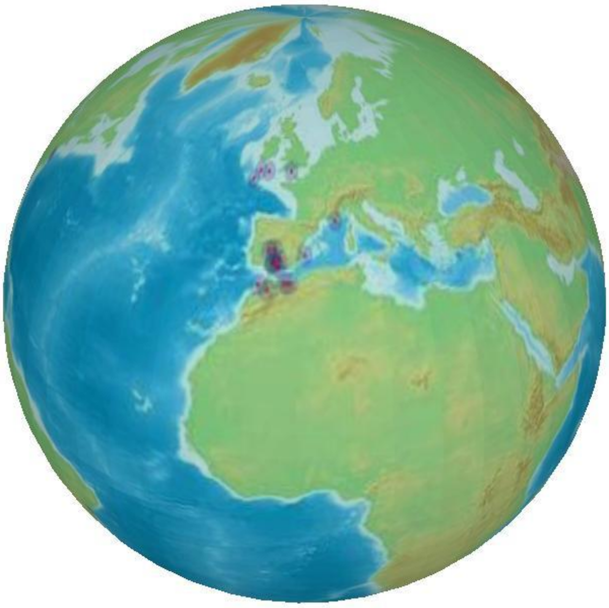

Figure 6. Visualization of places extracted, corresponding to Europe, for the newspaper ABC.

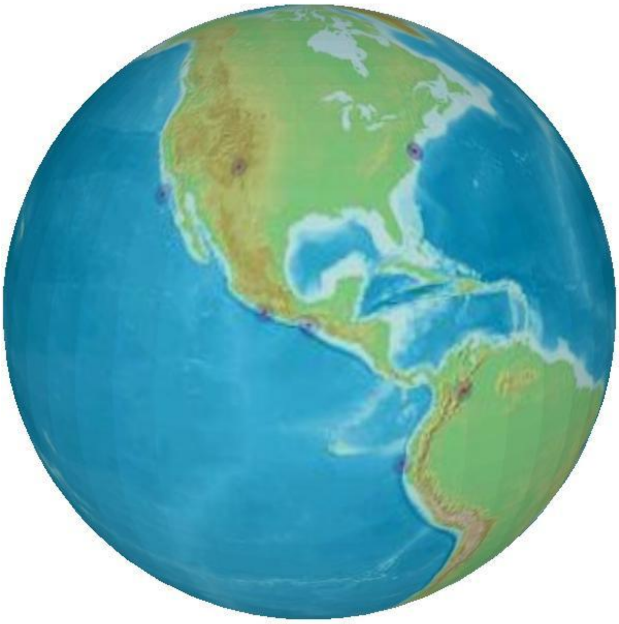

Figure 7. Visualization of places extracted, corresponding to America, for the newspaper ABC.

The interest of the results, presented here, is the capacity to automatically extract places from unstructured texts in order to visualize them and provide other kind of services to the users. In this way, we are working in order to provide to the users with the capability to manipulate textual travel narratives and newspapers by clicking places of interest or even having the capacity to visualize places according to time.

Results also show that despite the problems and difficulties inherent to services like the Geoparsing Web Service applied to unstructured texts, like the newspapers, it is possible to obtain meaningful information from dynamic web sources.

In this way, our article explores the possibilities given by a Geoparsing Web Service in order to extract and contextualize unstructured text documents. In our work we present the analysis and extraction of two cases 1) travel narratives and 2) online newspapers. In both cases, they are mapped into an Earth projection in order to visualize the marked places. Further work will be in the next axes:

- Identify other non structured documents to verify and check the Geoparsing extraction like historical documents or collections of documents like Wikipedia [9].
- Compare different Geoparsing methods (Metacarta, NetOWL, GeoLocator) in order to identify the best extraction process with the different applications.
- Having the extracted spatial position of the travel narrative or news we will have to find new ways of extract some spatial knowledge like GeoProfiling and find new geo visualization tools.
- The contextualization and disambiguation of the places named in each unstructured document sometimes it is not very clear so a work to contextualize and extract pertinent information of the original document with the use of ontologies is required [10]. Also, the use of information extracted from a large encyclopedic collection and the Web could be a solution to explore [11].
- Compare concepts, coming from different documents, such as the hierarchical nature of geographic space and the topological relationships between geographic objects in order to represent relationships between different documents [12].
- Geo-temporal criteria are important for filtering, grouping and prioritizing information resources [13]. In this way, the capacity to automatically link travel stories and newspapers on a time scale like the proposal done in articles of the Wikipedia shown by Bhole et al. [14] is an approach to be explored.
- Classify documents according to their implicit location relevance [15].

References

[1] S. Supavetch, S. Chunithipaisan, “The SQL-Based Geospatial Web Processing Service”. *International Journal of Computer Information Systems and Industrial Management Applications.* ISSN: 2150-7988. Volume 3 (2011). pp. 119-126. MIR Labs.

[2] G. de la Cruz M., F. Gamboa R., “Using User Interaction to Model User Comprehension on the Web Navigation.” *International Journal of Computer Information Systems and Industrial Management Applications.* ISSN: 2150-7988. Volume 3 (2011). pp. 878-885.

[3] R. R. Larson, “Geographic information retrieval and spatial browsing.” *Smith, L., Gluck, M. (eds.) University of Illinois GIS and Libraries: Patrons, Maps and Spatial Information,* pp. 81–124. 1996.

[4] B. Martins, and P. Calado, “Learning to rank for geographic information retrieval.” *GIR’10: Proceedings of the 6th Workshop on Geographic Information Retrieval.* ACM. ISBN: 978-1-60558-826-1. 2010.

[5] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, “Web-a-where: geotagging web content.” SIGIR '04: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in*

information retrieval, pp. 273-280, New York, NY, USA, ACM. 2004.

[6] R. Tobin, C. Grover, K. Byrne, J. Reid, and J. Walsh, “Evaluation of Georeferencing.” *GIR’10: Proceedings of the 6th Workshop on Geographic Information Retrieval.* ACM. ISBN: 978-1-60558-826-1. 2010.

[7] T. Qin, R. Xiao, L. Fang, X. Xie, and L. Zhang, “An efficient location extraction algorithm by leveraging web contextual information.” *GIS’10: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems.* ACM. ISBN: 978-1-4503-0428-3. 2010.

[8] I. Anastácio, B. Martins, and P. Calado, “Using the geographic scopes of web documents for contextual advertising.” *GIR’10: Proceedings of the 6th Workshop on Geographic Information Retrieval.* ACM. ISBN: 978-1-60558-826-1. 2010.

[9] J. Witmer, J., Kalita, “Extracting Geospatial Entities from Wikipedia.” *IEEE International Conference on Semantic Computing.* 2009.

[10] A. Zubizarreta, P. Fuente, J. M. Cantera; M. Arias J. Cabrero, G. García, C. Llamas and J. Vegas, “Extracting Geographic Context from the Web: GeoReferencing in MyMoSe.” *Proceedings of the 31th European Conference on IR Research on Advances in information Retrieval.* 2009.

[11] S. Cucerzan, “Large-scale named entity disambiguation based on Wikipedia data.” *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 708-716. Prague, June 2007.

[12] N. R. Brisaboa, M. R., Luaces, A. S., Places and D. Seco, “Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index.” *Geoinformatica* 14, 3 (Jul. 2010), pp. 307-331. 2010.

[13] B. Martins, H. Manguinhas, J. Borbinha, “Extracting and Exploring the Geo-Temporal Semantics of Textual Resources.” *IEEE ICSC*, pp. 1–9. 2008.

[14] A. Bhole, B. Fortuna, M. Grobelnik, D. Mladenic, “Extracting Named Entities and Relating Them over Time Based on Wikipedia.” *Informática*, 31(2007) pp. 463-468. 2007.

[15] I. Anastacio; B. Martins; P. Calado, “Classifying Documents According to Location Relevance.” *Proceedings of the 14th Portuguese Conference on Artificial intelligence: Progress in Artificial intelligence*. Lecture Notes in Computer Science. pp. 598-609. 2009.

Author Biographies

Rocío Abascal-Mena is a Professor at the Information Technology Department at the Universidad Autónoma Metropolitana – Cuajimalpa at Mexico city. She received his PhD in Computer Science at the Institut National des Sciences Appliquées de Lyon, France in 2005. Her research interests include Digital Libraries, Semantic Web, Ontologies, Information retrieval in ustructured texts, Human Computer Interaction, Natural Language Processing and Social Media.

Erick López-Ornelas is a Professor at the Information Technology Department at the Universidad Autónoma Metropolitana – Cuajimalpa in Mexico city. He received his PhD in Computer Science at the University Paul Sabatier in Toulouse France in 2005. His research interests include Geographical Information Systems, Geographical Visualization; Location based services, Geoweb and Remote Sensing imagery. In these areas he already published various papers on different international conferences. He also works on Human Computer Interaction applications and Context Awareness Systems.

J. Sergio Zepeda received the M.Sc. and Ph. D. degrees in Electrical Engineering specialty on Computing from Center for Research and Advanced Studies of the National Polytechnic Institute (Mexico) in 2003 and 2009, respectively. Currently, he is a Professor-Researcher of the Information Technology Department from the Universidad Autónoma Metropolitana – Cuajimalpa at Mexico city. His research interests include Web Engineering, Information Retrieval, Rich Interaction, Human-Computer Interaction, Usability and Interface Design. He developed a Semantic Interaction Model based on Web for database exploration on the Microbial Database of the Mexican Microorganisms Culture Collection. Interactive Applications with new paradigms for business, scientific and biological data are also of interest in his recent researches.