*GIR*Pharma:
A Geographic Information Retrieval Approach to Locate Pharmacies on Duty

Francisco M. Rangel Pardo
R&D Director
Corex Soluciones Informáticas
Grupo Fivasa / TIC Sinergies
Valencia (Spain)
francisco.rangel@corex.es

Loli Rangel Pardo
Linguistic Researcher
Corex Soluciones Informáticas
Grupo Fivasa / TIC Sinergies
Valencia (Spain)
loli.rangel@corex.es

Davide Buscaldi
Natural Language Engineering Lab
ELiRF Research Group
DSIC-UPV
dbuscaldi@dsic.upv.es

Paolo Rosso
Natural Language Engineering Lab
ELiRF Research Group
DSIC-UPV
prosso@dsic.upv.es

ABSTRACT

This paper describes an approximation based on geographic information retrieval with the purpose to give some solutions to the problem of searching pharmacies on duty in the Spanish territory. It is a novel investigation, which requires collaboration between multidisciplinary teams and that is beginning to show the first progress.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search and Retrieval, I.2.7.[**Natural Language Processing**]: Text analysis, J.3.[**Life and Medical Sciences**]: Health

General Terms

Design, Experimentation.

Keywords

Geographic Information Retrieval (GIR), Geographic Information Systems (GIS), Toponym Disambiguation

1. INTRODUCTION

Many ideas originate from necessity and curiosity, like the one we present in this work. The solution of some of the issues identified in this work came from the collaboration between diverse disciplines, in particular geographic information retrieval and geographical information systems, which played a major role as the "binding forces" that kept together the pieces of the developed system.

Everyone knows how much information exist nowadays in the big cloud that we call Internet, especially information that is constituted by unstructured but high-valued data. An example of this kind of data is constituted by the directions and timetables of the pharmacies that exist around us; in the case of this work, in Spain.

In Spain we have, right now, a strict regulation of the pharmacies, where it is dictated from the quantity of them there have to be at the most per habitants, to its localization and timetables and duties they have to execute.

Therefore, it is necessary for anyone needing to buy medicines at special times as at night, holidays or similar, to know pharmacies on duty in their neighbourhood, to go towards the best for him/her.

But all this information about pharmacies on duty is not available to the public through an exclusive site in Internet, but it is distributed by the diverse colleges of pharmacists in the diverse Spanish autonomous communities and provinces. In fact, in the level of towns, this information should be published by the own local webs, so the individual who needs to buy a medicine at a special time, has to know this, to know where he/she is, and to know where to go to search for this information, with the difficulty that this could imply to people without facility to use technology such us elders, foreigners and out of place people.

Our project is focused especially on inventorying every web site that publish information about pharmacies on duty, executing a geographic and temporal information retrieval about them on the geographical field where the individual would be and obtaining the georeferenced mentioned pharmacies in a friendly graphical interface, making easier to guide the individual to the nearest pharmacy.

Although we have been working on this project for a little period, we already have identified the tasks, some research problems and the methods to resolve them, and we have achieved some progress that may result of interest for the community.

In the next sections we briefly describe the project in all its parts: datawarehouse, crawling, indexing, geographic information retrieval, georeferencing, and geographic information visualization.

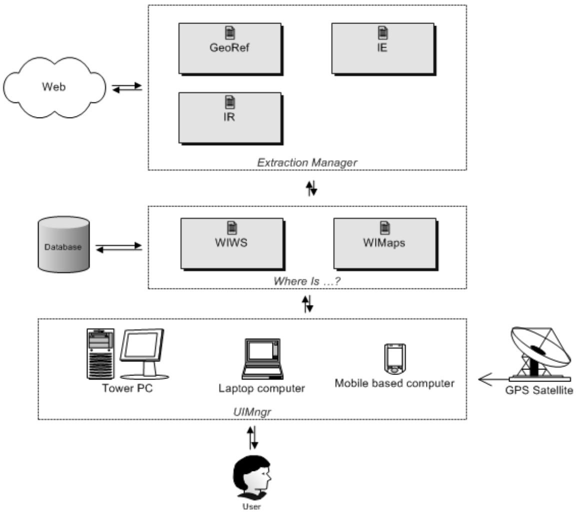
COM.Geo 2010, June 21–23, 2010, Washington, DC, USA.
Copyright 2010 ACM: 978-1-4503-0031-5…$10.00.

2. PROJECT DESCRIPTION

The objective of the project is ambitious and the range so wide, that the project has been divided into inter-dependent tasks that allow carrying out the research and the development in an increasing manner, at the same time that allow the interaction between the diverse teams, but with some autonomy in the achievement of the own objectives.

2.1 Architecture

The general architecture of the system is illustrated in the figure below:



Each one of the modules has been separated to help the researchers to work independently; they will be explained through consecutive points.

2.2 Datawarehouse Building

In information retrieval tasks, being spatial or not, data are usually scattered through a multitude of Web sites, as it is shown in [10], and in the case of the project we are describing, the mentioned information is, as we have already commented before, in the set of web pages maintained by the pharmacists colleges or in the pages maintained by local administrations.

The first step consists in the identification of this collection of pages and the building of a datawarehouse with the specified information for the extraction task.

The task of pages identification is done taking into account what we know about territorial organization of pharmacies in Spain. This means, we start in an autonomous community level, going down to the provincial level and from there on to the capital and to the level corresponding to the different towns.

During the construction of the datawarehouse, the information incorporated does not include only from the web addresses where there are data about the pharmacies, but also contextual information about the web, since contents are published in very different ways from one page to another.

The contextual information refers to the type of pages that is being published and how to access them (e.g. a static page with information about dates/hours and pharmacies localization, or a dynamic page, which has to be called in a determined way (GET/POST/SOAP) with determined parameters so that it would return the list with opened pharmacies), or the type of publication of the information, (e.g. html, pdf or even maps), in a way that the process of crawling had enough information to obtain the content in a satisfactory way.

As it can be appreciated, the work of construction of the datawarehouse comes into a tedious work if we try to incorporate at once the content of 17 autonomous communities, 50 provinces and 2 autonomous cities, and 8114 municipalities that Spain contains. So the work is done in a way that allows the successive incorporation of localities (at whatever territorial level) without affecting to the automatic integration with the rest of the system, starting from some sites of example with different territorial level and different model of publication of the contents.

2.3 Crawling and Indexing

The crawling process is based on the obtained information for each of the pages of the datawarehouse, that is, a selective crawler is built, which is able to obtain the html that contains the information that has to be extracted according to the type of page.

That means that a municipal page that publishes the list of dates and the pharmacies on duty at those dates is obtained completely for allowing the process of extraction of information to collect it and extract the interesting entities (it is described in Section 2.4). But a dynamic page that requires the user to introduce the date and the location, requires that the crawler executes a request of the type the page requires (normally get/post, but in some cases soap), with the suitable parameters to allow the page to reply with the list of pharmacies at this date and this location, and then the crawler obtains this final page to go on the following module.

The main problems that affect the crawling process are the variety of types of pages and the difficulty to access some of them. These issues complicate significantly the implementation of the crawler, which does not merely limit to the selective extraction of pages, but incorporates complex business logic, too.

2.4 Geographic Information Retrieval

The collection of textual data obtained from the previous points is used to extract the geographic and temporal information needed to determine the set of pharmacies that satisfy the user needs (that is, which are on duty at the moment and place required by the user).

The work at this stage consists in the identification of named entities such as dates, hours, locations, postcodes, etc., and the correspondent disambiguation of names to determine the precise manner to retrieve the information required.

One of the main tasks of disambiguation consists in determining the correct address from named entities such as street names. Disambiguating street names is not a trivial task, as it has been detailed in works such as [3] or [11] ; in [2] it has been observed that this kind of names have a higher ambiguity degree than other kinds of geographic names. They calculated the probability of a street name of being ambiguous in a collection of news documents is 0.83, while the same probability calculated for other kinds of geographic names is 0.58. The ambiguity of street names is particularly difficult to tackle, because of some issues that derive from the historical roots of their names. For instance, “plaza del Ayuntamiento” is a common name in Spain because every city has its own ``ayuntamiento” (city council).

Some road names do not change when they cross different administrative boundaries or when they belong to different towns: for instance, “Avinguda Dos de Maig” is a street that belongs both to Aldaia and Alaquas, two towns on the south of the city of Valencia, and the street itself marks a boundary between the two towns.

In other cases, roads keep the name of the city they are pointing to. Sometimes, it is not easy to understand a place name without knowing the source where the name has been extracted from. For instance, a “pharmacy in Aragon Avenue” may mean a pharmacy in Aragon Avenue, Barcelona or a pharmacy in Aragon Avenue, Valencia. If we know that this information has been extracted from a boundary near to Valencia (e.g. by GPS), then it is highly probable that the right referent is the place closest to this city [4].

Another non trivial tasks is the detection of acronyms and other words that do not give information to a georeferencing process, as for example:

C/ Manuel Pastor 18 *barriada el Carmen* valencia

(Manuel Pastor Street number 18 *in El Carmen area*, Valencia)

To help to the previous tasks, ontologies and dictionaries of named entities are used, and as it is observed in [5] they produce significant improvements at the time of disambiguate geographic elements.

Another aspect to take into account is the identification of temporal entities that can refer to explicit ranks as for example:

02/28/2010 from 10:00 pm to 08:00 am

Or to implicit ranks, as it is shown below:

Sunday, 28th of February of 2010

When the previous entities, both temporal and spatial, are identified and disambiguated, the relation between both of them is determined. Therefore, temporal and localized ranks are assigned to determine what type of information is required by the user from his/her localization to the requested moment.

One of the most common metrics at this type of problems, as it is shown by [8] for textual matching, is the distance measured in the number of words that separate them (“word distance”). So it is more feasible that a determined date would associate to a location that would be nearer to less words in distance than to other that would be at a farer distance. But here there is a problem of sequence identification too, as for example in the following list extracted from [6]:

02/28/2010 MAJOR PHARMACY

Major Street number 79, Alaquas.

03/07/2010 EAST PHARMACY

Iglesia Street number 10, Aldaia

The elements in italics have a lower word distance than elements that actually have to be matched, as the sequence in this case is firstly at a temporal entity and then spatially, and not on the contrary as it has been supposed in the example.

2.5 Georeferencing

Georeferencing here consists in assigning a geographical position (obtaining their geographic coordinates) to the information extracted in previous processes. Concretely, the collection of pharmacies located as on duty in the spatial and temporal field determined by the needs of the user.

Once the identification and the disambiguation of geographical entities have been carried out, the georenferencing should not represent a major issue, provided that databases with the coordinates of the physical directions of the pharmacies exist, but unfortunately this is not always true.

At this point we work on three levels, on the basis of dictionaries such as those mentioned for the disambiguation in Section 2.4.

- Local (municipality) data bases
- Cadastre data bases
- Google georeferencing API

The most up-to-date information is the one provided by municipal databases, which generally relate their street maps with parcels, and so with the territory. These data bases contain municipal information that is not always actualized by the Cadastre, like for example streets that disappear when a park is built or new streets emerge from the segregation of a higher street that is cut by a roundabout. A problem is that the access to those municipal data bases is not always publicly and freely available. Therefore, in these cases it is used the second option, that is the services of Cadastre are accessed.

Cadastre provides a web service that is able to return the coordinates of whatever postal direction under the Spanish territory. This georeferencer avoids the ambiguity requiring in a univocal manner province, municipality, acronym, street and number. The main problem is that it does not admit linguistic variations like omission of articles and/or prepositions, the morphological variety of the acronyms, like the diverse uses done of them (sometimes an avenue is named like street or vice versa), or even idiomatic variations like can be given in bilingual territories as Valencia, Mallorca, Cataluña, País Vasco (Basque Country) or Galicia. Therefore, it is necessary to obtain the bi-univocal correspondence between the geographical entities obtained from Internet, and the named entities indexed in the data base of the Cadastre.

At last, cases where search results fruitless due to the described problems, the Google georeferencer is used, although in some cases it works in a special way (e.g. randomly the search of Valencia can return *Valencia: and industrial city in Northern Venezuela*, or *Valencia: a city in Eastern Spain on the Mediterranean*)[5], this is overcome by determining the geographic environment where the research is done [9].

As well as the number of postal directions (including street and number) at a national level, although finite, is elevated and has difficult access without the massive usage of services like Cadastre (OVC), the number of pharmacies existing in Spain is limited and not as high as it could be expected (20.941 according to [6]). Therefore, part of the efforts of the research in the point of georeferencing is for the identification of the localization of pharmacies and its geographical link, such as geotagging described on [1]. Due to that, we move from using three types of services of georeferencing described above, to use an own index with the address of the pharmacies.

2.6 Optimum route calculation

Once the geographical point of reference of the individual is given, either being introduced by hand, by localization over a map, or by the reading of its GPS coordinates from the mobile device (as well as the collection of the pharmacies on duty, identified and georeferenced around it), it is of great interest to indicate the user what is the optimum way to reach each of them, either under his/her own decision, or under recommendation. Diverse technological institutes we are collaborating with, make use of technologies of last generation base on genetic algorithms, evolutional techniques or algorithms based on swarm intelligence [7].

2.7 Geographic Information Visualization

Once the user, the pharmacies on duty around him/her, and the way to reach them have been identified, it is necessary to provide the user with a system for visualization. There is a great quantity of geographical information systems on desktop as well the web, but on the mobile devices these systems are limited by the problems that have their operation in the five existing operative systems, like in the multitude of devices that operate particularly over them. At this point we have been working for several years to carry out products to the visualization of cartography and geographical elements, and we have a Geographical Information System called "Where Is…?"[1] in a Web platform and in a Windows Mobile one. Moreover, we collaborate with technological institutes that have been acrying out advanced research in this field in the rest of mobile devices (Symbian, Android, iPhone and Blackberry), in order to offer a solution for the highest number of users.

3. CONCLUSIONS AND FUTURE WORK

We have been working on a novel, interesting and really ambitious project, that incorporates different technologies and disciplines, where the guiding and unifying thread is the match of geographic information systems and the geographic information retrieval.

We have described the tasks, its problematic and shown the state where we are now and the path we still have to cover.

The work to be done includes to increase the information of the datawarehouse, to incorporate more intelligence to the process of crawling, to improve the disambiguation and georeferencing, among others.

One of the tasks that is being investigated like an alternative to the current georeferencing model, as we commented in Section 2.5, is georeferencing every existing pharmacies in Spain, (20.941 on December of 2007 according to [6]), and disambiguate in the geographical field that surrounds the user (as described in [4]). We estimate that the complexity of georeferencing should decrease and the quality of overall results increase.

4. ACKNOWLEDGMENTS

The work of F.M. Rangel and M.D. Rangel was funded by R&D department from Corex Soluciones Informáticas, SL.

The work of Davide Buscaldi and Paolo Rosso was partially funded by the TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) research project

5. REFERENCES

[1] Amitay, E., Har'el, N., Sivan, R., Soffer, A. Web-a-where: geotagging web content. SIGIR'04. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004. Pp. 273-280

[2] Buscaldi, D., Magnini, B: Grounding Toponyms in an Italian Local News Corpus. Proceedings of GIR'10 workshop, 18-19th Feb. 2010, Zurich, Switzerland

[3] Buscaldi, D., Rosso, P.: A Conceptual Density-based Approach for the Disambiguation of Toponyms. International Journal of Geographical Information Systems, 2008 Pp. 301-313

[4] Buscaldi, D., Rosso, P. Map-based vs. knowledge-based toponym disambiguation. Workshop on Geographic Information Retrieval. Proceedings on the 2nd international workshop on Geographic Information Retrieval, 2008. Pp. 19-22

[5] Buscaldi, D., Rosso, P.: Using GeoWordNet for Geographical Information Retrieval. In Revised Selected Papers CLEF-2008, Springer-Verlag, LNCS(5706), Pp. 863-866

[6] Consejo General de Colegios Oficiales de Farmacéuticos. La Distribución de la Farmacia en España. Contestación al Informe Ecorys, May, 2008.

[7] Dorigo, M., Gambardella, L.M. Ant colonies for the traveling salesman problem. Biosystems, 1977. Pp. 73-81

[8] Mascardi, V., Locoro, A., Rosso, P. Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation. IEEE Transactions on Knowledge and Data Engineering, June 2009. IEEE Computer Societe Digital Library. DOI: 10.1109/TKDE.2009.154

[9] Min Duc, T., Nishimura, T. Geographic Information Retrieval System Using Semantic Relationships Between Multiple Layers. Proceedings of the 10th International Conference on Information Integration and Web-based Application and Services,2008. Pp. 305-311

[10] Turton, I., Jaiswal, A., Gahegan, M. Geographic Information Retrieval from Disparate Data Sources. Penssylvania State University. Geocomputation, 2007

[11] You-Heng Hu, Linlin Ge. A Supervised Machine Learning Approach to Toponym Disambiguation. Advanced Information and Knowledge Processing the Geospatial Web. Springer London, 2007. Pp. 117-12

[1] http://www.where-is.es in http://world.corex.es