Where the streets have no name – Experiences in GIR for a developing country

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

ABSTRACT

This paper gives a short overview of a project on Geographic Information Retrieval in developing countries in the form of an experience report based in Honduras. It provides some insights into encountered challenges of resource discovery, and georeferencing due to low Web coverage and informal location references as well as tested or proposed solutions to address them, including search via alternative means such as social networks.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Design, Documentation

Keywords

Geographic Information Retrieval, Local search

1. INTRODUCTION

This paper reports on an extended research stay in Honduras, a developing country in Central America. It gave the opportunity to experience firsthand the challenges of digital divide and technology use in developing countries in the context of researching working solutions for actual industry partners. It challenged many previously held assumptions of the author regarding availability, granularity, quality, and detail of data, as well as cultural aspects. The initial project idea was presented in [1], while [3] gives a more extensive retrospect. In this instance, nameless streets were only a small part of a large amount of interesting challenges encountered. Some examples of location references along with their translation are presented in the following:

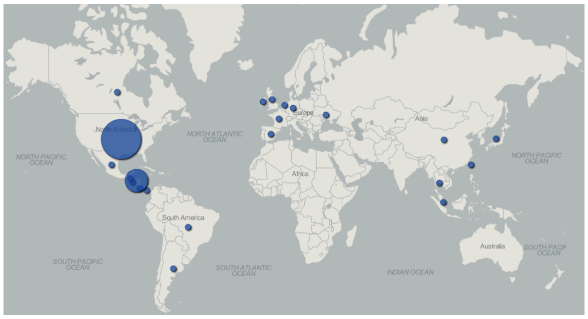

Figure 1: Mapped hosting countries for .hn domains

- *Casa Morazán, Centro de Tegucigalpa, contigua a la calle peatonal, frente al antiguo Cine Variedades (ahora tienda de electrodomésticos Elektra) - casa color rosado/rojo.* The building with the name Morazán, in the center of Tegucigalpa, in the continuation of the pedestrian street, opposite the former cinema (which is now a home appliances shop), house has pink/red color.

- *En el anillo.* This location is on the ring road (about 30km in length) that circles the capital Tegucigalpa, which is not mentioned in the reference.

- *Colonia Kennedy 2da Entrada, Tegucigalpa.* Near the second entrance to the neighborhood Kennedy.

- *3a Calle, Tegucigalpa Honduras.* 3rd street (running east to west) in Tegucigalpa, which means it is in the old center.

With such references, many current GIR approaches will still work, as they focus more on the neighborhood or city levels. However, it decreases the chance of successfully and precisely geocoding an actual entity. Additionally, they still require the availability of reasonable amounts of documents, gazetteer data, or geocoders to work. Unfortunately, other major issues were a low Web coverage and generally underdeveloped Web infrastructure. We illustrate some GIR-specific challenges, based on our outset of a development process [5]:

Assessing assumptions To understand the situation and the use of local search, the information needs, and the preferred sources, we performed a user study [6]. It showed among others that people prefer word-of-mouth, use Facebook as a main online source, and have issues with the imprecise addressing scheme.

Resource discovery We identified various semi-structured data sources apart from the general Web that we wanted to crawl that potentially carry relevant location information. The goal was to access them either by API calls or with Deep Web crawlers [11]. User-generated

GIR'13 November 05 - 08 2013, Orlando, FL, USA
Copyright 2013 ACM 978-1-4503-2241-6/13/11
http://dx.doi.org/10.1145/2533888.2533937 ...$15.00.

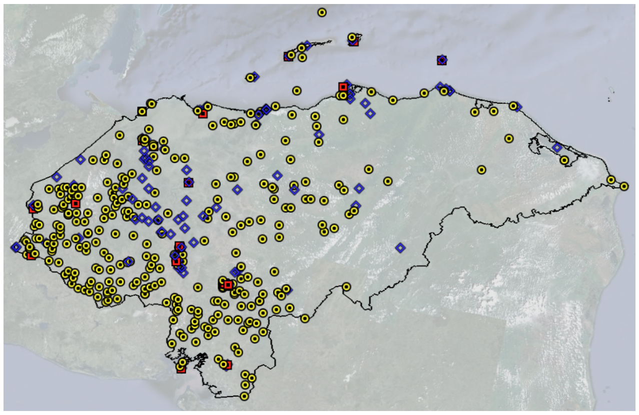

Figure 2: Mapping of geotagged Wikipedia articles, yellow ⊙: merged articles, red ⊡: Spanish, blue ◇: English

content could in part also help to build the index [13]. For Web crawling, we wanted to examine the digital divide hypothesis [12] and found strong indicators by the amount of Honduran domains hosted outside the country (Fig. 1, only 26% are hosted inside the country, 60% in the US; even 25% of government domains are hosted in the US [7]).

Wikipedia integration Interestingly, the Spanish-speaking country is covered by more English georeferenced Wikipedia articles than Spanish ones. To include both languages, we developed a translation approach which could unify most articles [4] (Fig. 2).

Geoparsing Old city centers from a colonial time often have a grid structure that allows easy addressing. Yet, for the capital, this only covers about 5% of the area, the rest consists of few named larger thoroughfares and smaller streets with unknown name so landmark navigation and neighborhood referencing is often used. This makes geoparsing very challenging, as can be seen from the previous examples.

Geocoding Due to the lack of exact addresses, geocoding sources mostly only cover broad granularities at the level of residential areas. For better reference identification, including landmarks from data sources such as OpenStreetMap is an ongoing process [2].

There is related work on various issues of this work, such as country-level Web search engines [10], SMS search in developing countries, [9] or using the Web as a spatial datasource in emergent countries [8].

2. FUTURE WORK

Not all issues could be solved during the research stay, yet a number of successes were possible as well as initial implementations for location-based services with a local mobile phone operator. Others are open issues for future work.

We have found that many businesses and public places do not set up Web pages anymore and instead use Facebook pages. As this is also the mainly used online source for local information, we propose to expand the Web crawling with social search based on Facebook information. While there are structured fields for locations, many pages use either a descriptive or very broad location reference as seen in the examples in the introduction.

In these cases, a potential location can easily be assumed. Actually grounding the descriptions is more difficult. The descriptions have to be understood and the constituent parts interpreted in the right order to construct the references and directions. An open challenge is that if high-level information is missing, the actual city and even the relation to Honduras has to be inferred from other features. In these cases, we cannot just use inverse focused crawling with a known list of names. We aim to use a combination of Web-level and social-centric graph measures which might help to derive a locality measure from surrounding entities in the graphs and thus assist in both crawling and grounding.

3. CONCLUSION

We have presented a short experience report of a project aiming to transfer GIR methods to a developing country. Of the numerous challenges, some could be addressed with a combination of existing and modified methods. Challenges due to imprecise addressing and a low availability of data remain, and fuel open research questions regarding multi-source geocoding and the use of social media for local search.

4. REFERENCES

[1] D. Ahlers. Towards Geospatial Search for Honduras. In *LACNEM 2011*, 2011.

[2] D. Ahlers. Multi-source conflating index construction for local search in a low-coverage country. In *LA-WEB 2012*, 2012.

[3] D. Ahlers. In search of Honduras – Case report of developing local search for a developing country. In *LWA 2013 – Lernen, Wissen, Adaption*, 2013.

[4] D. Ahlers. Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles. In *ECIR2013*, 2013.

[5] D. Ahlers. Towards a development process for geospatial information retrieval and search. WWW'13. 2013.

[6] D. Ahlers and N. Henze. ¿Donde está? – Surveying Local Search in Honduras. In *MWB2012 – Workshop on Mobility and Web Behavior at MobileHCI2012*, 2012.

[7] D. Ahlers, J. Matute, I. Martinez, and C. Kumar. Mapping the Web resources of a developing country. In *GI Zeitgeist 2012*, 2012.

[8] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and C. A. Davis, Jr. Discovering Geographic Locations in Web Pages Using Urban Addresses. In *GIR'07*, 2007.

[9] J. Chen, B. Linn, and L. Subramanian. SMS-based contextual web search. In *MobiHeld'09*, 2009.

[10] M. Mendoza, H. Guerrero, and J. Farias. Inquiro.CL: a New Search Engine in Chile. In *WWW'09*, 2009.

[11] D. Mundluru and X. Xia. Experiences in Crawling Deep Web in the Context of Local Search. In *GIR'08*, 2008.

[12] K. T. Nakahira, T. Hoshino, and Y. Mikami. Geographic Locations of Web Servers under African Domains. In *WWW'06*, 2006.

[13] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR'07*, 2007.