REVISTA COLOMBIANA DE COMPUTACIÓN
Volumen 13, número 2
Págs. 37- 52

Multi-source entity-based conflation for local search

Dirk Ahlers[1] ‡

Fecha de recibido:19/06/2012 Fecha de Aprobación: 20/10/2012

Abstract

Local search is a well-established mode of Web search engines today that makes georeferenced entities available for search. For most developed countries, a huge amount of data is directly available on the Web and can be extracted and processed by search engines. Results are mapped with great precision and often are correct down to the granularity of the correct building. In many developing countries, the Web coverage is much lower, only little information is directly available, and the granularity may be much worse, both in the stated addresses or locations on Web pages and in the source data used in the geocoders. To develop a local search that still can provide sufficient precision, a hybrid mode of index construction has to be followed that identifies and integrates other sources of geospatial information to increase the coverage. The location part can be exceptionally difficult, as imprecise addresses and low-granularity geocoders do not allow precise coordinates to be used in mapping. We therefore propose an approach that is designed to increase coverage and precision by collecting and merging entity data from multiple sources. Its purpose is to combine the strengths of individual data source and reduce the impact of their weaknesses. We use the example of the Latin American country of Honduras, to describe the approach and potential data sources.

Keywords: *Web Search, index construction, geospatial search, geocoding, conflation, entity search, Honduras.*

[1] NTNU – Norwegian University of Science and Technology Trondheim, Norway, dirk.ahlers@idi.ntnu.no.

1. Introduction

The use of location-based search engines with a huge amount of searchable data and extensive, reliable results is taken for granted today in many parts of the world [1]. These are built on top of the major search engines' indexes of the Web and use specialized entity extraction techniques to identify location and addresses [2] as well as entity names of businesses, restaurants, museums, people, parks etc. and additional information such as opening hours, phone numbers, and more [3]. This data is taken directly from the Web in the form of Web pages. Additional information is used in the extraction process to ground place descriptions and verify found entities [4], [5], [6]. This is aided by two additional types of data sources. For the geospatial aspect, gazetteers are geographical thesauri that contain place names and the relation between them and are used to identify place names. Geocoders are then used to translate broad place names or exact addresses to coordinates to pinpoint onto a map. Secondly, external data sources such as yellow page data are used to verify found entities. Additionally, cross-referencing between multiply-encountered entities can aid the verification.

An interesting challenge occurs in developing countries that have a much lower Web coverage and are not very well represented in POI databases or structured yellow-page-style directories. For a geospatial search engine we are developing in Honduras [7], we cannot rely on many of the characteristics taken for granted in the developed world. We have to adapt to the rather poor data situation at hand. The search engine has to include specific other data sources that can deliver additional entities or features or additional granularity to found places. We therefore aim to improve geocoding granularity by applying entity resolution to geospatial data sources. For example, this allows pinpointing entities whose location is only roughly described in, e.g., as a location on a Web page, to be improved by merging it with basic information from a more exact POI database. While the former contains more textual information, the latter may only name the title, but have more accurate geographical information available. In combination, this can mutually improve the available data.

In the remainder of the paper, we will discuss challenges for a local search engine, discuss some data sources and then present our approach at data source selection, entity extraction, geocoding, entity merging and conflation, and index generation.

2. Challenges and Background

Due to our work in Honduras, we use the country as a representative example. However, the challenges discussed here also apply to many other developing countries with similar characteristics [8].

In terms of the viability of local Web search, we have identified two main characteristics of Honduras that distinguish it from other, more industrialized countries that are usually the focus of the literature about geospatial search. The first is the generally low information density on the Web, which means that we simply cannot expect the high amount of information on the Web that is available in other countries The second is a rather imprecise addressing system, which means that textual representations of locations or addresses often can not capture the position with sufficiently high granularity, because, e.g., house numbers are not used, some streets are nameless, directions are only given broadly or in relation to landmarks [7], [9].

2.1. Honduran Web Coverage

Honduras has a very low Web coverage with less then 1500 domestic .hn domains, of which even a majority are hosted out of the country. Even when including generic .com domains, which are often used due to their easier setup, we estimate only around 20000 domains [10]. As only about 11% of the population in Honduras has access to the Internet, this is also a sort of chicken-egg-problem, which might improve in the future. However, there are other ways to improve the situation now already and make use of the available information as good as possible. One way is by also including other sources apart from Web pages. We will therefore follow a hybrid approach of both Web search with georeferencing of documents and additional access to specific data sources, which may hold more precise features that can be combined.

2.1.1. User Behavior

For example, current patterns of user practices only sparingly use local Web search, but rely very heavily on word-of-mouth or social networks. This is of course dependent on the perceived quality and ease of use. In a user study [9] we asked participants about their mode and sources of information gathering. The answers are shown in Fig. 1 and demonstrate that Web search on its own is by far not as strong as in other countries, but that social media, especially Facebook, is a preferred source. Conversely, we might therefore expect a lot of relevant information in this source.

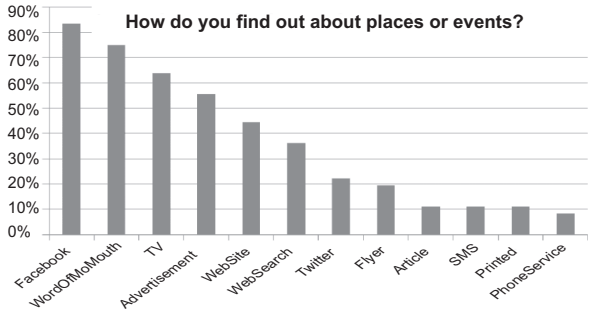

Fig. 1. Percentage of respondents that use modes of local information gathering (multiple answers possible).

2.1.2. Web Search

For the basic discovery and processing of Web data, we will use methods of focused Web crawling [11]. The seeds are generated by search engine sampling or inverse focused crawling [12] to make use of the advantage of big search engines in indexing the Web and even remote generic domains that carry Honduran content. Additionally, we are also aiming to better understand and trace the Honduran domain and IP space [10]. In addition to the Web, we will follow a multi-source approach and also specifically target Deep Web sources [13], [14]. Entity extraction is used to identify locations and entity names [15], [16]. A following geospatial data fusion step is based on entity reconciliation to merge the same entities found in different sources [3]. We extend this location-based entity reconciliation to a conflation approach to improve positional accuracy as discussed in Section 3.

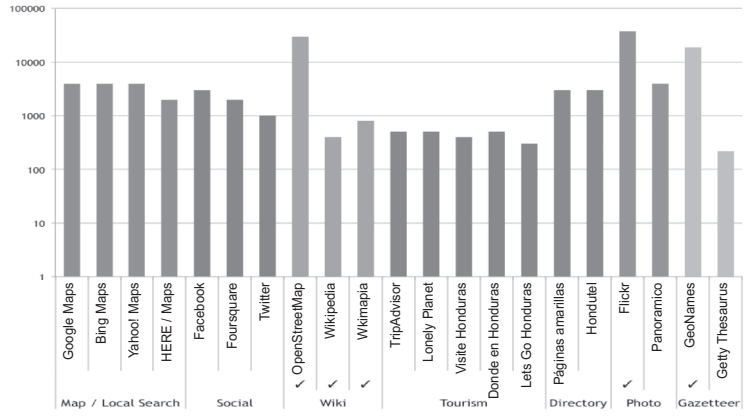

Fig. 2. Data source overview grouped by type: initial estimate.

2.2. Structured Data Sources

A huge number of structured or semi-structured data sources such as Facebook, foursquare, OpenStreetMap, Wikipedia, or GeoNames have been identified that may provide data and be a target for additional entity retrieval [7]. The records inside the sources can range from simple POIS with a name and a pair of coordinates to rich descriptive information with supporting information on certain features. In most cases, the richness of individual entries and the overall amount of entries show an inverse relation.

Fig. 2 shows an initial estimate on the amount of data available. Due to the widely varying numbers, it is plotted at a logarithmic scale. It includes data sources from the areas of map/local search, social media, wiki-style sources, tourism, directory, and gazetteers. This is only a very broad estimate on the order of magnitude to direct resources towards the most promising sources. More accurate methods [17], [18], [19] may be used for better estimates. The figure does not report on sources with fewer than 200 estimated entries for the country, to keep the graph manageable. General search engines are also left out at this point, as they do not contain actual coordinates. With the risk of overreporting, these contain an estimate of 1-3 mio pages. The pages are of course to be considered, but the final search engine may instead crawl itself for resource discovery [11]. The number can thus also be understood as the amount of Web pages to crawl.

Two requirements have to be fulfilled for the use of a data source in the search engine. The first is obviously the amount and quality of the data, but the second is the availability of said data. Licenses and permissions for API calls to local search engines often either forbid automatic extraction of data or set limits on their use. While our intent is not to build a competing service, the restrictions apply. On the other hand, some data is rather difficult to search and extract from the sources. Without going into details, data sources with easy access are marked with a checkmark in the graph.

Geonames was found to be the most comprehensive gazetteer with a good coverage throughout the country. A gazetteer is a geographic thesaurus that contains place names at varying granularities, usually down to a city or city district level together with geographic relations, population data and, most important in this case, geographic locations [20]. It can additionally contain information about places of interest, geological formations, etc. and is used as a geographic knowledge base and to ground place names to coordinates.

In working to access the datasources, we have further developed a method to manage overlap between Wikipedia articles from different

languages. With a translation and merging approach we can reach a more comprehensive coverage [21]. Photos from Flickr are out of scope of this article. It can be noticed from the graph that OpenStreetMap (OSM) is also estimated to contain a very high amount of data. We will come back to this source later.

2.3. Coarse Addressing of Places

Lacking a formal, high-granularity addressing system in most of the country, Honduran location references usually have no house-number or building-exact address. In some cases, they might even just reference a large street or boulevard or the broad neighborhood or city district. These then lack the high granularity needed for exact geocoding and pinpointing them on a map. In old areas of the capital or in smaller cities a rectangular street grid of *calles* and *avenidas* exists, which then allows for block addressing, but this only covers a relatively small space, around 5% in the case of the capital.

In most regions, location references are given by city name, city district and sometimes the street name. Various other forms of descriptions have evolved that allow finding a certain building. Often these are given additional directional information such as nearby landmarks or well-known buildings. Sometimes a description is accompanied by a sketched map, a so-called croquis to help with orientation. Some examples of such addresses found on the Web are "*Zona Jacaleapa, frente a Colonia Honduras, Tegucigalpa*" (a part of the city next to a neighborhood), "*final de Bulevar Morazán*" (the end of a 2.5km street), "*3a Calle, Tegucigalpa Honduras*" (a street in the old city), "*En el Anillo*" (somewhere on the ring road circling the city). This poses some challenges to address and place recognition, toponym resolution, and finally geocoding.

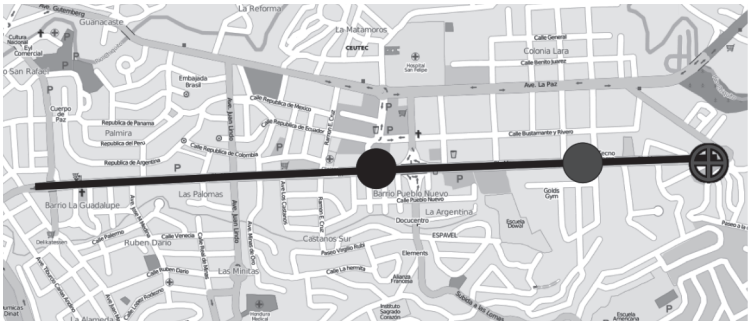

Fig. 3. Different result options for geocoding Tigo "al final de Bulevar Morazán". The boulevard is shown as a blue line, its centroid as a red circle, the centroid of its last third as an orange circle, and the actual position as a green crosshair to the east.

The goal is that a business with a coarse location reference such as "*al final de Bulevar Morazán*", which references one end of a 2.5km boulevard, should be geocoded to an exact coordinate and not to a makeshift of just the centroid of the street or the last 500-1000m as seen in Fig. 3. To reach a sufficiently high accuracy, the system has to improve upon the geocoders by using additional data.

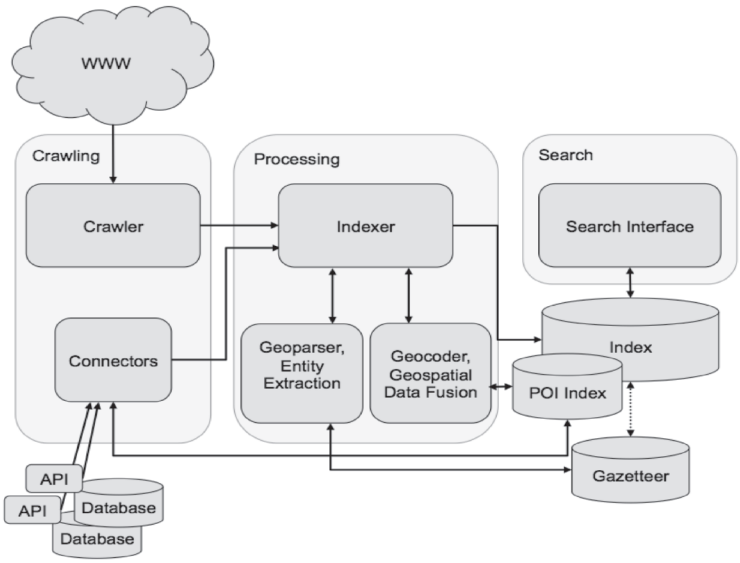

Fig. 4. Architecture of the overall search engine.

3. Approach

The index of a local search engine is a hybrid index that combines a textual index with the geospatial aspect of coordinates as latitude and longitude. This then actually allows for geospatial search by enabling proximity queries, results within a certain region, etc. It might also include an index of structured entities that is linked to the textual and geospatial index. The basic architecture is shown in Fig. 4, already including adaptations to extract information from structured or semi-structured data sources via API calls, connectors, domain-specific parsers, extractors, and extensions in the index structure.

3.1. Hybrid entity-based geocoding

As discussed before, the geospatial granularity of textual location references can be rather low. Our aim is to still be able to geocode found entities with a high reliability to only a small radius of uncertainty. Low-

granularity location references can be extracted and geoparsed from documents [22], [23]. However, simply geocoding them would result in a coordinate with a high uncertainty [24]. In some cases, geocoding a large entity such as a city would best result not only in a coordinate, but also in an annotated extent of the entity or maybe even in an outline. However, some commercial geocoders can return the granularity level as a feature type. This enables to distinguish between city, neighborhood, street, or building accuracy. However, it only notes the granularity, but does not improve it.

We propose an approach of entity-based geocoding conflation that simultaneously extends entity resolution systems. The idea is based on the observation that there are two distinct types of data sources that can complement each other. On the one hand, we have Web pages with a high amount and depth of information, yet often only very coarse location references. On the other hand, we have point-based location databases that may have only very rudimentary information about entities, often not more than the name, but possess very accurate coordinates for these. If it becomes possible to merge these two types of sources, the resulting entities would be vastly improved.

The concept is depicted in Fig. 5. Multiple entities go in, with different information depth and different location accuracy. After the merging, the mutual best information is combined and attached. The example shows how this would work for a Web page with a broad address and a POI of only the building name and its position.

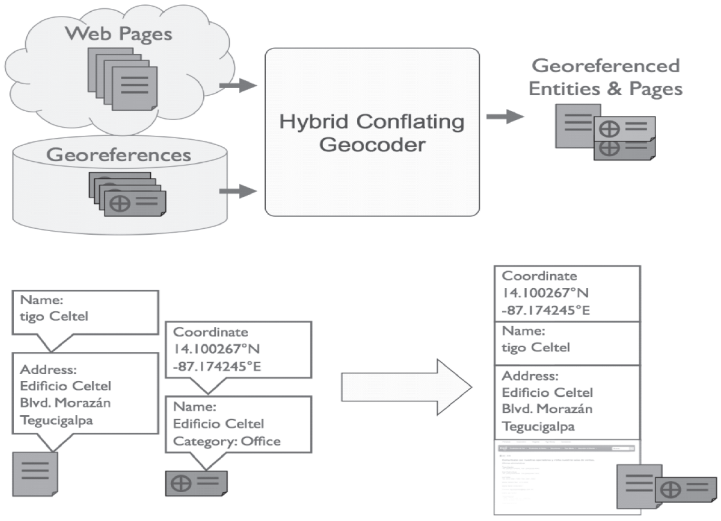

Fig. 5. Entity conflation of textual data with named point data; example.

This proposal differs from purely textual entity reconciliation in that it also needs to address the geospatial dimension, which can of course also be uncertain and imprecise [25]. A similar approach to use additional information to improve geocoding has before been explored by, e.g., [26] who use tax records to better estimate the size of parcels and thus correct geocoding results based on interpolated TIGER lines[2]; or [27], who use a multi-stage geocoding approach with multiple geocoder databases. Our approach differs in that we use more general, freely available data without country-dependent availability and that we address a very different level of granularity where parcel identifiers or house numbers for disambiguation are not available. Therefore the conflating geocoding cannot be based on textual address alone, but needs to consider other features such as a roughly similar location and a matching entity name. This means that it is limited to places that contain a named entity, which aligns with the original goal.

A good example for a data source with very high positional accuracy is OpenStreetMap (OSM)[3] which also was identified above as containing a huge amount of relevant data. OSM is a great repository of information, but it is not yet very accessible to users searching for specific places, which might explain its relative low-key publicity. However, OSM has evolved a lot from just providing maps of the road network to actually containing a large amount of georeferenced, named, and sometimes numbered buildings and the named businesses that reside in them.

We therefore aim to use OpenStreetMap as an external geocoder database for point-based location data of named entities. This slightly extends the geocoder concept, as geocoders already use additional information about road networks or house number patterns to extend gazetteer data. However, they usually reach their limits when we talk about a region where house numbers are uncommon. In Honduras, while house numbers theoretically exist, they are used very seldom to locate a place.

In the process, if an entity that was found on a Web page or in another source could only be a roughly geocoded or only a broad location could be extracted by geoparsing, we use the name, broad location, and available other characteristics to query it in OpenStreetMap. Additional features such as classifications, can be taken, but are often unreliable or imprecise. The conflation needs to take certain name variations into account, as well as integrate a distance measure to streets, landmarks, or

[1] A shapefile format. For geocoding, it contains the start end end house numbers along a stretch of road, therefore numbers in between have to be interpolated. http://www.census.gov/geo/maps-data/data/tiger.html
[2] http://www.openstreetmap.org/
[3] http://www.openstreetmap.org/

other POIs, as OSM data often has points on the map without relation to a street. The matching street for a point then needs to be found by, e.g., reverse geocoding.

It is also possible to exploit additional hints in the textual location about the side of the road, or spatial relations to other buildings, such as ”near to”, ”after”, ”behind”, or ”opposite of”. However, this requires an iterative approach that might revisit one entity after other referenced entities are grounded.

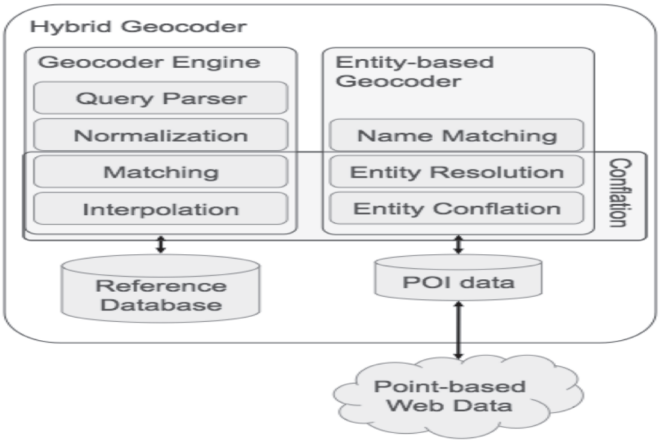

Fig. 6. Conceptual architecture of the geocoding conflation system.

Fig. 6 shows the concept of the geocoder component. On the left side are the usual components of a geocoder, the right side shows the proposed additions to allow it to work on named entities and improve the spatial accuracy.

Fig. 7. Entity example from OpenStreetMap and Google Maps.

For example, the German embassy has two ways to express its address. The first, *Avenida República Dominicana # 925, Callejón Siria, Col. Lomas del Guijarro, Tegucigalpa,* is the official address. In this case, all streets have a name and even house numbers are used. However, this would not be used by anyone trying to find it, and the more common description would be *Avenida República Dominicana, subiendo 2 cuadras del Ministerio Público, en una pequeña calle sín salida a lado derecha*[4]. This means that it is 2 blocks above (which in this case means a topographic above, up the hill) the Public Prosecutor's office in a side street. In the textual example, *Avenida República Dominicana* could be found, but it is a long street that winds up the hill. However, we could also find *Ministerio Público* to narrow down the area. This would mean an iterative approach to first ground and verify buildings used as landmark and then work from them onwards. Malls, churches, large embassies, gas stations, prominent fast food shops, schools etc. can serve the landmark purpose. Armed with this additional information, we could check that area in OpenStreetMap and actually find the *Embajada de Alemania* there. As Fig. 7 shows, this location is very accurate. In this specific area of town, it is also better than Google Maps, which locates the embassy in a different part of town and does not yet have the actual street in its database. The hint to both the street and the prosecutor's office would even help to determine that the Google result is too far away from the area in question. In another example (cf. Fig. 8), a restaurant only gives its location as *Blvd. Morazán.* It does not show up in Google Maps, but in a Web search. Its foursquare location is slightly off, and its own Facebook page only gives the street name. In a yellow page, it is described as opposite a nearby mall. In OSM, it is properly positioned, which means that the correct positioning of the named entity would be available after entity resolution and conflation.

Fig.8 . Entity examples from Facebook, yellow pages, foursquare, and OpenStreetMap, counter-clockwise

[4] http://www.tegucigalpa.diplo.de/Vertretung/tegucigalpa/de/02/Lageplan/Lageplan.html

The approach proposed here has so far been evaluated in a proof-of-concept stage on selected entities; work on a larger scale is yet to be done. We expect to encounter some limitations as we broaden the scope of included entities. The spatial relations can be difficult to understand because they may not only indicate position such as opposite or nearby, but also may use topographical relations such as above, below, behind, which will only be considered as nearness. The notion of the end and beginning of a street is difficult to obtain and exists as a colloquialism separately for each street Inverse geocoding to find the street where an entity is located can lead to ambiguous results. Small name variations are detectable, but if names differ too much between the sources, identification is seriously hindered. While in theory, alternate names can be annotated in OSM, the field was almost never used. Furthermore, multiple entities with similar names may exist in such a way that the information density is insufficient to identify or separate them. Changed information is usually rather quickly reflected in the OSM data, but can still linger on for a long time on the Web Finally, a huge amount of locations is simply not yet available in the sources, so that some entities will only be present in one source so that the approach obviously will fall short.

4. Conclusion

We have proposed the concept for an entity-oriented Web search and entity reconciliation and conflation approach. It is designed to exploit multiple sources of the Deep Web to improve the results of a local search engine that can only access the surface Web. The main point for the use of additional sources is to improve the geospatial location accuracy for the geocoding stage within the search engine. The hybrid geoparsing/geocoding/entity reconciliation system is designed to take rough textual descriptions from Web pages with high information density and, by using the results of named entity extraction, lookup these entities in spatially more precise sources to actually pinpoint these on the map.

At the moment, the concept is a work-in-progress that is being tested on a small manually selected testset. For some cases, the entity resolution and conflation works easily, others are the focus of ongoing investigations to enable or improve it. Limitations include complex directions that need more sophisticated patterns, big mismatches between names used in different sources that cannot be identified, or incorrect, imprecise, or missing data in the mapping services as well as issues in reverse geocoding an entry to the correct street in all cases and limited access to public geocoding services for the region. We further envision to include additional crowdsourcing sources as well as social

networks into the described processes. The main issue there will be the limited possibility to freely search Facebook, foursquare, or other services and to actually match entities with incomplete locations.

While the work described in this paper is aimed towards Honduras, we expect the system to also be applicable to other countries with similar characteristics. The hybrid geocoder could then be used to improve coverage and location accuracy in any region where high-quality point-based geospatial data is available to improve upon low-granularity places found on the Web.

Acknowledgement

This work was done while the author was at UNITEC – Universidad Tecnológica Centroamericana in Tegucigalpa, Honduras.

References

[1] D. Ahlers, “Local Web Search Examined,” in Web Search Engine Research, ser. Library and Information Science, D. Lewandowski, Ed. Emerald, 2012, vol. 4, pp. 47–78.

[2] D. Ahlers and S. Boll, “Retrieving Address-based Locations from the Web,” in GIR '08: Proceedings of the 5th International Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2008.

[3] D. Ahlers, “Business Entity Retrieval and Data Provision for Yellow Pages by Local Search,” in Integrating IR technologies for Professional Search Workshop @ ECIR2013, 2013.

[4] R. S. Purves, P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, and B. Yang, “The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the internet,” International Journal of Geographical Information Science, vol. 21, no. 7, pp. 717–745, 2007.

[5] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger, “Design and Implementation of a Geographic Search Engine,” in WebDB 2005, A. Doan, F. Neven, R. McCann, and G. J. Bex, Eds., Baltimore, Maryland, USA, 2005, pp. 19–24.

[6] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, A. S. D. Silva, and J. Clodoveu A. Davis, "The Web as a Data Source for Spatial Databases," in Anais do V Brazilian Symposium on Geoinformatics, 2003.

[7] D. Ahlers, "Towards Geospatial Search for Honduras," in Proceedings of the Latinamerican Conference on Networked and Electronic Media LACNEM 2011. San José, Costa Rica: Universidad Latina Costa Rica, 2011.

[8] C. Farvacque-Vitkovic, L. Godin, H. Leroux, F. Verdet, and R. Chavez, "Street Addressing and the Management of Cities," The World Bank, Washington, DC, USA, Tech. Rep., 2005.

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

[10] D. Ahlers, J. Matute, I. Martinez, and C. Kumar, "Mapping the Web resources of a developing country," in GI Zeitgeist 2012, Proceedings of the Young Researchers Forum on Geographic Information Science, ser. IfGi prints, vol. 44. AKA, 2012, pp. 117–122.

[11] D. Ahlers and S. Boll, "Adaptive Geospatially Focused Crawling," in CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2009, pp. 445–454.

[12] "Location-based Web search," in The Geospatial Web. How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society, A. Scharl and K. Tochterman, Eds. London: Springer, 2007.

[13] D. Mundluru and X. Xia, "Experiences in Crawling Deep Web in the Context of Local Search," in GIR '08: Proceedings of the 5th International Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2008, pp. 35–42.

[14] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the Deep Web," Communications of the ACM, vol. 50, no. 5, pp. 94–101, 2007.

[15] M. J. Cafarella, J. Madhavan, and A. Halevy, "Web-Scale Extraction of Structured Data," SIGMOD Rec., vol. 37, no. 4, pp. 55–61, 2008.

[16] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A Survey of Web Information Extraction Systems," IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 1411–1428, 2006.

[17] A. Anagnostopoulos, A. Z. Broder, and D. Carmel, "Sampling Search-Engine Results," in WWW '05: Proceedings of the 14th international conference on World Wide Web. New York, NY, USA: ACM, 2005, pp. 245–256.

[18] A. Dobra and S. E. Fienberg, "How Large Is the World Wide Web?" in Web Dynamics: Adapting to Change in Content, Size, Topology and Use, M. Levene and A. Poulovassilis, Eds. Springer, 2004, pp. 23–44.

[19] K. Bharat and A. Broder, "A technique for measuring the relative size and overlap of public Web search engines," Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 379–388, 1998.

[20] L. L. Hill, "Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints," in ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. London, UK: Springer, 2000, pp. 280–290.

[21] D. Ahlers, "Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles," in ECIR2013, 2013, short paper.

[22] S. E. Overell and S. M. Rüger, "Identifying and grounding descriptions of places," in Proceedings of the 3rd ACM Workshop on Geographic Information Retrieval, GIR 2006, R. Purves and C. Jones, Eds. Seattle, WA, USA: Department of Geography, University of Zurich, 2006.

[23] J. Leidner, Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal Publishers, 2008, PhD thesis.

[24] D. Ahlers and S. Boll, "On the Accuracy of Online Geocoders," in Geoinformatik 2009, Osnabrück, ser. ifgiprints, W. Reinhardt, A. Krüger, and M. Ehlers, Eds., vol. 35, Münster, 2009, pp. 85–91.

[25] V. Sehgal, L. Getoor, and P. D. Viechnicki, "Entity Resolution in Geospatial Data Integration," in Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, ser. GIS '06. New York, NY, USA: ACM, 2006, pp. 83–90.

[26] R. Bakshi, C. A. Knoblock, and S. Thakkar, “Exploiting Online Sources to Accurately Geocode Addresses,” in GIS '04: Proceedings of the 12th annual ACM international workshop on Geographic information systems. New York, NY, USA: ACM Press, 2004, pp. 194–203.

[27] G. Lovasi, J. Weiss, R. Hoskins, E. Whitsel, K. Rice, C. Erickson, and B. Psaty, “Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree?” International Journal of Health Geographics, vol. 6, no. 1, p. 12, 2007.