Mapping the Web resources of a developing country

Dirk AHLERS [a,1], José MATUTE [a], Isaac MARTINEZ [a] and Chandan KUMAR [b]
[a] *Unitec – Universidad Tecnológica Centroamericana,Tegucigalpa, Honduras*
[b] *OFFIS – Institute for Information Technology*

Abstract.
To develop a search engine tailored to a specific country, its relevant Web resources have to be identified. The first step in this resource discovery is to retrieve Web documents that contain content related to the country. In our work, we aim towards a search engine for Honduras, a developing country in Latin America with an emerging use of the Web. In a preparatory step, we aim to find relevant domains prior to setting up our own crawler so that they can serve as seeds to the crawler, which can then be run in a more focused way to retrieve the actual Web pages. We initially use two angles for domain resource discovery. The first is to use the data available from the DMOZ catalogue, the second is to use major commercial search engines for an overview of Honduran domain names used on the Web. We report on initial results as well as the used methods and additionally give data about the geographic distribution of Honduran Web servers both inside and outside the country.

Keywords. Geospatial Search, Geospatial mapping, Resource Discovery, Crawling, Network infrastructure, Domain names

1. Introduction

In an ongoing project, we develop a geospatial Web search engine for the Latinamerican country of Honduras [1]. Besides structured databases for basic information such as gazetteers, directories, or named entities, the search engine should mainly target the Web, in this specific case, only the Honduran Web. One of the multiple challenges is the resource discovery process by a Web crawler which traverses the Web link graph in a search for relevant pages. This is of course the main feature of a crawler [2] which allows it to reach most parts of the visible Web. Obviously, the crawler needs some initial URLs as entry points into the Web graph, the so-called seeds. Theoretically, it is possible to only start with very few domains. However, if we want to restrict the crawler to only pages from a single country, it would be harder to keep the crawler focused on the relevant pages without downloading huge parts of the Web that are not of interest. Focused crawling has been developed as a technique based on heuristics and machine learning to steer a crawler towards relevant pages for general [3,4] as well as for geospatial topics

[1]Corresponding Author: Dirk Ahlers, Unitec – Universidad Tecnológica Centroamericana, Sistemas Computationales, Facultad de Ingeniería, Tegucigalpa, Honduras, ahlers@dhere.de

[5]. Yet a large seed set is preferable because it keeps the focus sustainable for longer durations.

This article details the steps taken towards resource discovery for the domain names and Web resources, aiming at a delineation of the Honduran Web. Similar work has been done already for Portugal [6,7], Argentina [8], and Chile [9]. We adapt and develop some of their methods to arrive at a better understanding of the Honduran Web.

1.1. Honduran domain name and IP space

The country-code top level domain (ccTLD) for Honduras is `.hn`. It uses structured second-level domains (.edu.hn, .gob.hn, etc.), but their use is not enforced and user-selected second-level domain names can be registered. As is common practice, the local domain registrar nic.hn does not give out the list of registered domain names. Therefore, other means of retrieving them are discussed in the following sections. However, some statistics were available. Figure 3 shows that there are about 5780 domain names registered in Honduras, distributed over the direct ccTLD and the second-level namespaces. Compared to the population of about 8 million this puts the domain ownership rate at only about 0.07%. Additionally, a large number of relevant domains are registered under general top level domains (gTLD) outside of the .hn ccTLD.

2. DMOZ directory data

A source for a first overview of Honduran Web information is DMOZ[2]. It is a popular seed source for Web crawlers (e.g., [10]). The majority of the country speaks only Spanish, but English is used along the Caribbean coast and islands. We therefore downloaded the English and Spanish language versions from its geographical hierarchy for Honduras. This resulted in 421 results for English[3], and 96 for Spanish[4]. This discrepancy already shows a particular bias.

We first examined whether the links were from the .hn TLD or at least contained this term somewhere in the URL. The results are shown in Figure 1. While there are much more English articles than Spanish ones, the Spanish ones prove to be more local. Inspection shows that the former contains many overview pages, travel information, and other general external information about the country (from, e.g., WHO, BBC, Wikipedia, universities), while the latter contains pages that are actually located within the country. Yet, for both, many pages exist from entities inside Honduras that have a generic TLD. One example is this university, which uses the domain unitec.edu.

Therefore, this step was insufficient and instead, a more thorough manual assessment of all links was performed. We classified the actual provenance of the information, either maintained by or being the own page of a Honduran entity; or from outside the country. This classification is plotted in Figure 2. Totals are slightly lower, as some pages were not reachable. Obviously, the use of the .hn TLD and the provenance of the information diverge heavily. While the simple check for .hn only gives 2,4% relevant links for English

[2] The Open Directory Project `http://www.dmoz.org/`

[3] `http://www.dmoz.org/Regional/Central_America/Honduras/`

[4] `http://www.dmoz.org/World/Espa\%C3\%B1ol/Regional/Am\%C3\%A9rica/Honduras/`

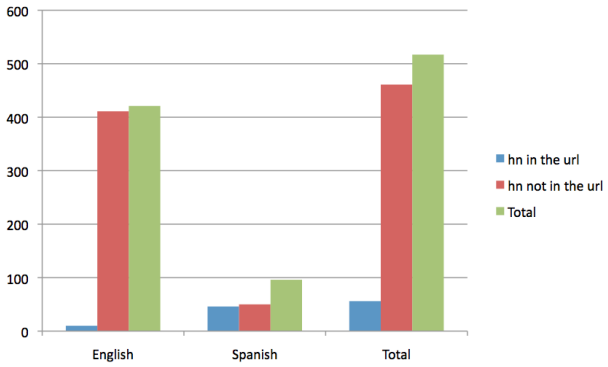

Figure 1. Name analysis of Honduras data in DMOZ

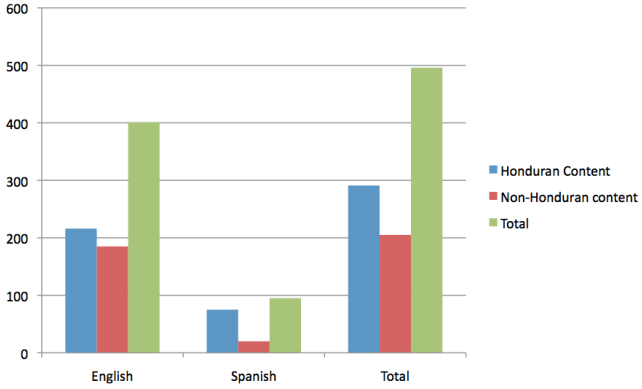

Figure 2. Provenance analysis of Honduras data in DMOZ

and 47,9% for Spanish, the provenance analysis puts these values at 53,9% and 78,9%, giving a more realistic estimate. It also reduces the apparent bias in the English results towards out-of-country sites and instead shows a strong underdevelopment of the Spanish version.

3. Search engine sampling

While we will use our own Web crawler, we have neither the resources nor the need to crawl the whole Web. We therefore use focused crawling to stay on relevant pages [5], which needs support by a good seed set.

A possibility to generate seeds besides using DMOZ is to use inverse focused crawling [11], which poses well-crafted queries to large search engines. While there is no direct support to search directly for domain names, both Google and Bing understand the "site: search" operator, which allows to specify part of the hostname, e.g., "site:.gob.hn" for governmental domains. Each search query will only return the top k results, with k at 1000 documents for both Google and Bing. The count usually includes multiple documents per domain. We therefore try to partition the result set with increasingly specific conjunctive queries if the result set for a query is larger than k=*1000* as a divide and

domain name	expected	found	percentage
.hn	4260	685	16%
.com.hn	1253	93	7%
.gob.hn	85	59	69%
.org.hn	84	29	34%
.net.hn	56		
.edu.hn	41	26	63%
.mil.hn	1	1	100%
sum	5780	893	15%

Figure 3. Assigned Honduran domain names count and retrieved domains

conquer strategy. For each of the .hn hierarchies, we initiate a site search. The terms used in the conjunctions are taken from a list that consists of all department and municipality names. Additionally, it includes typical descriptions of entities taken from a dictionary.

The retrieved pages were reduced to their domain name, duplicates were removed, and only the first named domain part was selected so that subdomains are excluded. While adding more terms to the query reduces the result set size, it can also exclude documents from the results. Overall, the results reported here have a slight bias towards well-ranked pages and the reported numbers will be underestimated. The overall count of found names as presented in Figure 3 is rather low. Two possible explanations are erroneous data for the expected values or a high number of registered domains that are not in active use. They are supported by the counts for .net.hn, where we did not find any of the supposedly existing domains. We therefore estimate that the 15% found domain names are a reliable lower bound with an upper bound at about 25%, setting the number of active domains at below 1500. Due to the insufficiency of only examining .hn domains, we will search the general Web for gTLDs containing relevant content in our future work.

4. Mapping the Honduran Web

Early work in geographic information retrieval has suggested that infrastructure location can in part be related to content location at least on a country-scale [13,14]. We therefore analyzed the hosting locations for Honduras, following other work examining developing countries [15].

We used a commercial service from ipaddressapi.com that is free for a limited amount of queries. Range lists were found to be too inexact or lacking coverage. The country assignment is shown in Figure 4 on a logarithmic scale. A large amount is actually assigned in Honduras, but the majority in the US, with other American countries following behind. The *no response* domains could not be resolved. A deeper inspection of the US hosts revealed these to be mostly in southern countries associated with a large Latinamerican population, while the other countries are often related to the owners or investors of businesses. However, we also found 25% of governmental domains hosted in the US, confirming the suspicion of a 'digital divide' [15].

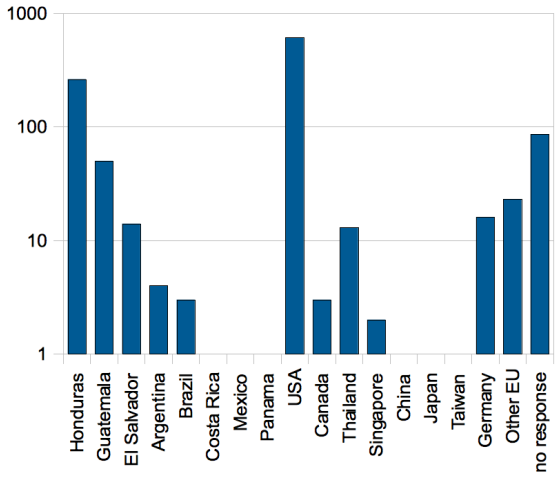

Figure 4. Distribution of hosting countries for .hn domains

5. Conclusion

Honduras presents interesting characteristics of its domain infrastructure. The amount of registered domain names in Figure 3 stands in contrast to the names that we were actually able to find. Analyzing the DMOZ domains, .hn domains only represent an estimate of between 5% to 61% of all relevant domains. Therefore, the search for out-of-country domains promises a wealth of additional resources. The hosting distribution shows that even for country-based information, foreign hosting often seems to be the preferred solution. The discovery of out-of-country domains is still ongoing. Because of these diverse issues, the discovery of the Honduras Web remains a challenge.

References

[1] 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.

[2] C. Olston and M. Najork, "Web Crawling," *Foundations and Trends in Information Retrieval*, vol. 4, no. 3, pp. 175–246, 2010.

[3] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," *Computer Networks*, vol. 31, no. 11-16, pp. 1623–1640, 1999.

[4] A. Micarelli and F. Gasparetti, "Adaptive Focused Crawling," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer, 2007, vol. 4321, pp. 231–262.

[5] 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.

[6] D. Gomes and M. J. Silva, "Characterizing a National Community Web," *ACM Transactions on Internet Technology*, vol. 5, no. 3, pp. 508–531, 2005.

[7] D. Gomes, A. Nogueira, J. Miranda, and M. Costa, "Introducing the Portuguese web archive initiative," in *Proceedings of the 8th International Web Archiving Workshop*, Aaarhus, Denmark, September 2008.

[8] G. Tolosa, F. Bordignon, R. Baeza-Yates, and C. Castillo, "Characterization of the Argentinian Web," *Cybermetrics*, vol. 11, no. 1, July 2007.

[9] M. Mendoza, H. Guerrero, and J. Farias, "Inquiro.CL: a New Search Engine in Chile," in *WWW '09: 18th International World Wide Web Conference (WWW in Ibero-America track)*, ser. WWW '09. ACM, 2009.

[10] P. Srinivasan, F. Menczer, and G. Pant, "A General Evaluation Framework for Topical Crawlers," *Information Retrieval*, vol. 8, pp. 417–447, 2004.

[11] D. Ahlers and S. Boll, "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.

[12] F. McCown and M. L. Nelson, "Agreeing to Disagree: Search Engines and Their Public Interfaces," in *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2007, pp. 309–318.

[13] M. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic Locality of IP Prefixes," in *Internet Measurement Conference (IMC) 2005*, Berkeley, CA, October 2005.

[14] K. S. McCurley, "Geospatial Mapping and Navigation of the Web," in *WWW '01: Proceedings of the 10th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2001, pp. 221–229.

[15] K. T. Nakahira, T. Hoshino, and Y. Mikami, "Geographic locations of web servers under african domains," in *Proceedings of the 15th international conference on World Wide Web*, ser. WWW '06. New York, NY, USA: ACM, 2006, pp. 989–990.