Towards a Development Process for Geospatial Information Retrieval and Search

Dirk Ahlers
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
ahlers@dhere.de

ABSTRACT

Geospatial search as a special type of vertical search has specific requirements and challenges. While the general principle of resource discovery, extraction, indexing, and search holds, geospatial search systems are tailored to the specific use case at hand with many individual adaptations. In this short overview, we aim to collect and organize the main organizing principles for the multitude of challenges and adaptations to be considered within the development process to work towards a more formal description.

Categories and Subject Descriptors

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

General Terms

Design, Documentation.

Keywords

Geospatial Web Search, Search engines, Information and Knowledge Management, Developing countries

1. INTRODUCTION

Location is an important organizing principle for many Web search tasks. In most industrialized nations the search for locations features prominently within search engines and users are used to seamlessly using local search with a multitude of correct results. It works this well because there is both good data available and tailored technology to make use of it. In many developing countries, the situation is gravely different. Local search may not be as accessible, important places are missing, place descriptions are lacking, geocoding may be hard, or the information density and depth is rather low. Errors or inaccuracies may further complicate the situation, if information is even available in the first place. Motivated by the case of geospatial search in developing regions, we want to abstract from specific country-level challenges for local search and carve out more general principles.

This is based on own previous work geospatial search engines for Germany [1, 3] and a research stay in Honduras [2], where we had to challenge many previously held assumptions and were forced to reevaluate our views. In trying to gather the necessary data, gazetteer information, Web pages, IP space structures etc., understand place names, directions, and location names, as well as develop the necessary approaches, we first rather informally started to try different ways and means to respond to the challenges. Similar research projects has been described for Chile [8], Portugal [6], Brazil [5], or Germany [7]. With the necessary hindsight, we now aim towards a more formal description of the development process for geospatial search and services. In this first step, we focus on overarching fields and principles of study even if they may not apply to all cases.

The large questions fall back on the well-known challenges of *where* to get the data, *what* to do with it, *how* to cope with shortcomings? Non-geospatial aspects as well as technical implementations are out of scope. The remainder of this paper will outline aspects of the proposed development process for a country-level geospatial search engine, focusing on requirements and feasibility analysis.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. DEVELOPMENT PROCESS OUTLINE

Depending on the case, approaches can range from using API calls of established local search engines to crawling the whole Web or the Deep Web or buying necessary data from specialized providers. However, even this decision is usually based on a deliberation of the different available options, which itself is further based on the availability or knowledge of these options and their characteristics. The following presents some of the most important points to consider. Main challenges include the requirements and data situation; and the analyzing, extracting and indexing of geospatial data.

We previously described the process life cycle of geospatial Web information in a crawling search engine [4]. We now extend and refine this process towards a more abstract and general development process for a geospatial search engine by discussing the main principles and fields of work.

Assessment and understanding A first step is to get an overview of what data sources, applications and services may be available and relevant. Specific undertakings range from market analysis, viability analysis, and data source investigations over requirements engineering to data analysis and user studies, regarding search or (mobile) applications. These can provide initial insight into user needs, the availability of data, and requirements for subsequent steps.

Enablers Quite an obvious step, a market analysis should find out if the intended solution is actually new and needed and especially if other participants can provide some basic services or data.

Users As a search engine is an offer towards users, their requirements and situation as well as social considerations have to be taken into account. Existing privacy and security concerns should also be evaluated and later addressed. Informal interviews, user studies, or usage observation can provide valuable insight.

Data The data situation is the most important, as it plays an important role in the feasibility analysis. This includes how much data is available, what its characteristics and quality is, which additional sources are available, as well as coverage, density, and distribution analysis.

Country-specific characteristics Locations tend to be described and used differently across the world. This includes availability and descriptions of addresses and directions, used granularity and addressing schemes, potential multi-language-use, levels of reliability, preferred media, size and distribution of domains, and other cases to consider. For example, a low Web coverage can be offset with a stronger use of structured data sources or user-generated content.

Building a gazetteer and knowledge base To aid in the extraction of geospatial data, a bootstrapping of known geographical placenames is normally used. Such gazetteer data can either be directly available or needs t be collected and combined from multiple sources.

Discovery and analysis of data sources To get a good overview, various data sources have to be explored to understand the type of information they offer and estimate the amount of data available.

Extraction and analysis of data Specific extraction methods have to be developed. These especially concern geoparsing, i.e., the identification and extraction of location references, and geocoding, i.e., the grounding of location references to geographic coordinates.

Web crawling For crawling a country-specific Web, the characteristics and boundaries of the country in the Web have to be defined and delineated. The actual crawling can then better focus on the relevant parts of the Web.

Source integration Local search is not just a document search but is also an entity search in the sense that it models georeferenced documents as well as the actual georeferenced entities described in the documents, following a hybrid search approach. For structured data and multiple Web results, cross-correlation and entity resolution across all results is needed.

Infrastructure and interfaces Finally, based upon the available data and potential augmentations from additional analysis steps, interfaces to actually search the index are developed, which carry their own set of considerations. This is of course based on an overall search engine infrastructure including storage, ranking, etc.

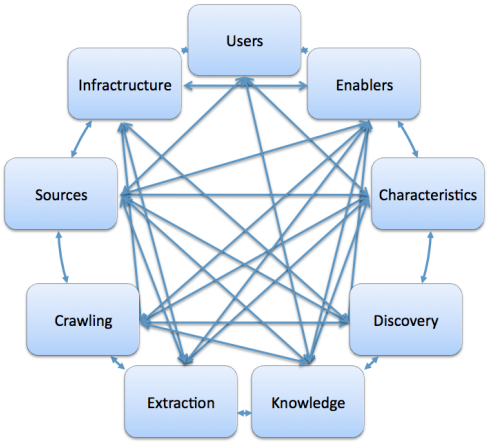

Figure 1: Main fields of consideration for geospatial search and retrieval development

3. CONCLUSION

We have presented a brief overview of issues to consider when building geospatial search systems. This was based on our own experiences as well as relevant related literature. Some issues are well-known in the literature, while others are of a more subjective, anecdotal nature and would need broader input for generalization. We hope to deepen this discussion in the future to arrive at a more formal description of the development process including and classifying all the relevant aspects, their dependencies, and cross-sectional issues which could only be touched briefly here. Of interest would also be the impact of different factors for different regions of the world, ranging from developing to industrialized countries.

4. REFERENCES

[1] D. Ahlers. *Geographically Focused Web Information Retrieval*, volume 18 of *Oldenburg Computer Science Series*. OlWIR, 2011.

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

[3] D. Ahlers. Business Entity Retrieval and Data Provision for Yellow Pages by Local Search. In *IRPS Workshop @ ECIR2013*, 2013.

[4] S. Boll and D. Ahlers. A Web more Geospatial: Insights into the Location Inside. In *WebEvolve2008 Workshop @ WWW08*, 2008. Web Science Research Initiative.

[5] 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.

[6] D. Gomes, A. Nogueira, J. Miranda, and M. Costa. Introducing the Portuguese web archive initiative. In *8th International Web Archiving Workshop*, 2008.

[7] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and Implementation of a Geographic Search Engine. In *WebDB 2005*, 2005.

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