Business Entity Retrieval and Data Provision for Yellow Pages by Local Search

Dirk Ahlers
Geo Search Consultant
Oldenburg, Germany
ahlers@dhere.de

Abstract

Yellow pages have for a long time been the major providers of business' addresses and contact information. Nowadays, many people use local search engines for such information needs. Contrary to Web search, results for local search are expected to be much more precise and actually describe a pinpointed entity instead of giving out long result lists. On the other hand, yellow page providers still exist and market their data and services. One aspect of this is that they also license their data as reliably, manually edited entities to local search providers. Therefore, when yellow pages want to use the Web as an additional source of information about their entities, using local search engines is not a viable option and a tailored search solution has to be developed.

1 Introduction

Most local search providers rely on classifieds or yellow page data as a starting point and enhance it with their own crawling, extraction, and aggregation [Ahl12]. In this paper, we will focus on the origin of the the yellow page data. In most cases, they are licensed from commercial providers that collect and publish relevant business information. For example, in Germany, businesses buy a presence in the yellow pages and have their data manually edited, curated, and verified by the yellow page providers. For this, the yellow pages rely on a surveying department of human editors who use the Web, and many other sources, for manual information gathering. The following describes a project that was undertaken with a yellow page provider in Germany to support and improve this process. The objective was to use the Web as a primary source of raw data to find out more about existing and prospective customers. This introduces a special type of professional search which is a combination of local search, aggregation, and analysis as a data preparation process. The editors cannot use the local search engines but have to search within the general web pages to find even small mentions and disaggregated mentions. This is an interesting constraint, as the system has to break out of the feedback loop by being autonomous and not relying on local directories and search engines that might just perpetuate their own data. Previously, major Web search engines such as Google or Yahoo were used to enrich the data, e.g., find out the homepage/domain for an entity or find/verify phone numbers, find entities at a certain address etc.

We therefore built a vertical search engine for object-level business information extraction. It acts as an additional source of semantically enriched data for the data survey department which until then had to rely on

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

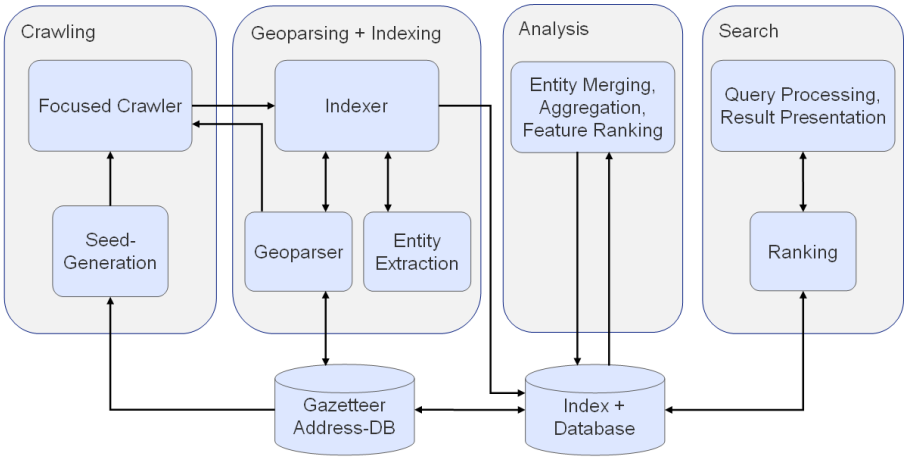

Figure 1: Architecture of the Entity Retrieval System

manually gathered data. The interest was in the location-based entities that can be found on the Web [Him05]. The pages themselves were of secondary interest only as a source for object extraction (c.f. [CKGS06, CMH08]), as they contained the entities and their business-related properties. These included the addresses, entity names, phone numbers, email addresses, commercial register entries and tax numbers, contact persons, homepage, line of business, products etc. The developed business-entity search solution was based on own previous work in geospatial Web search engines and data extraction [AB07]. It realizes the central concepts of focused resource discovery and extraction of location information by a geoparser. It follows a high-granularity approach of extracting location references down to the level of individual addresses and buildings [AB08]. This is an optimal foundation for the yellow pages, as they are also based on addresses at the house number level. The motivation for the business entity search was twofold. First, to semi-automatically enrich existing directory entries and second, to discover new promising candidates. An emergent requirement was to distinguish between different entities, perform classifications and clusterings, assign reliability scores and also to capture potential changes and deviations of entities over time. Additionally, the system mandates high levels of recall and the ability to work through results in an ordered fashion to make sure editors would be able to process all found entities.

To ease this process and classify cases based on which action, if any, was necessary, the system had to match directly into the provider's database and link with existing entries. Furthermore, it has to integrate into the client's workflow at the survey department. Due to reasons of data quality, all actual changes in the yellow pages database itself are still performed manually. However, the preparatory steps of discovering, extracting, verifying, and clustering the information have been automated using information retrieval and extraction techniques. This ensures that the information base, the key property of the business, remains of high quality.

2 Approach

We have developed a general architecture for local search as presented in [AB07, Ahl11], which consists of the major components of crawling, extraction, and search interface. Fig. 1 shows the adaptations implemented for this specialized system. The main change concerns a new component for extensive analysis of data in the index, respectively the database used for the extracted entities. It performs the enrichment, duplicate detection, merging, and aggregation of entities based on similarity analysis. The indexer was extended with components for the entity extraction and there is a tighter integration of gazetteer and index.

2.1 Resource discovery

We use a geospatially focused crawler for resource discovery to capture a large part of northern Germany in accordance with the area served by the particular yellow pages provider out of the larger network. Similar work at country-level crawling has been done for a variety of regions and countries [MBS05, GS08, MGF09, BYCMR05]. Additionally, the system can be set up to selectively only cover certain areas, for example if only a limited region has to be updated for a special edition or to go through the whole covered area region by region.

We initiate the system with links from DMOZ as seeds for the crawler. As part of the large and multi-source information the yellow page provider own, they have specific address data available for their whole served region. We integrate the relevant parts of the address database into a gazetteer, a geographical thesaurus, to extend it down to the address level, which will also be used in subsequent extraction steps. Based on the part of the gazetteer that is activated for a crawl, the seeds are selected by choosing those parts of the DMOZ geographical hierarchy that match the cities and districts selected for a crawl. From these initial links, the focused crawler traverses the Web in an adaptive best-first approach [AB09]. This ensures that pages with addresses from the relevant regions are actively sought out and prioritized.

2.2 Extraction

For each Web document retrieved by the crawler, its unstructured contents are analyzed to identify and extract geographic entities and their attributes. The high-granularity dictionary-based approach of our previously developed address-based geoparser [AB08] is applicable to Web pages that contain precise addresses within a city. The geoparser allows for multiple locations per page that are linked at a page level to the document representation in the index. Entries are merged on an inter-page level by later steps. This is especially necessary for business listings or lists of branch offices that have multiple addresses on a page.

In a subsequent step, the environment of an address on a page as well as the page as a whole is examined for additional information about entities located at the address. An entity name extractor aims to derive the company, person, or organization that is referenced by an identified address. Several other parsers are implemented to subsequently derive additional business information such as entity names, phone numbers, email addresses, commercial register entries and tax numbers, contact persons, homepage, or keywords above from the Web pages. This follows work in the extraction of structured data, e.g., [CKGS06, CMH08]. Using external knowledge in the form of whois data is not feasible due to the restrictions for German domains that disallow automatic processing.

2.3 Analysis

To consolidate entities and to prepare the crawled and extracted results for an entity-based search, the system uses a two-stage approach. In a first step, entities are extracted from individual Web pages as described above and stored. After a crawl, or at reasonable intervals, the individual entities that may occur on multiple pages and multiple domains are merged. This entity fusion aims to merge entities and combine attributes found on multiple separate sources. Thus, it abstracts from individual Web pages and provides a business entity view. All URLs that contributed to a merged entity are added as sources to the attributed that were found on them to trace the origin of data snippets. The merging and enrichment of entities is supported by external domain knowledge sources for a better extraction and identification process [LKS08]. This includes already known information such as business directory entries of the yellow page provider itself, legal form designations for company names, product name databases, or regional phone code coverage.

The entity fusion again is a two-step approach. It is based upon the individual properties of an entity. The initial similarity of entities is determined by a shared address. Since the address' spelling is normalized, no ambiguity on, e.g., street name, can occur. The entity name only comes second, as we see a lot of deviations on this. These can often be resolved by small editing distance, reordering of words, disambiguation of abbreviations, etc. The trivial cases are those where all properties of two instances agree. They are identified and the source list of the merged entity properties receives the sources of both entities. Thresholds on the similarity scores for properties and overall entities determine how to process conflicting information or outliers. Depending on the similarity measure over all properties, entities may be merged, annotated with less reliable alternative names or properties, or only grouped for later manual inspection. For the various properties, multiple similarity scores as well as ranking and grouping functions are defined. For textual content such as company names, string similarity measures such as editing distance are adapted by, e.g., incorporating company legal type designations. More difficult cases are, for example, the inclusion and exclusion of the owner's name in the business name. In cases where we could extract the owner's name independently, we can identify these and mark it as an unclear naming, otherwise, they are marked as potentially related. Phone numbers enable other heuristics for ranking, such as a matching of the area code to the address (cf. [AHSS04]) or a preference for numbers indicating a central switchboard instead of an extension. Frequency counts on the sources, adjusted by the type of source, such as same or other domain, business listing site and similar measures are used to estimate the official domain of an

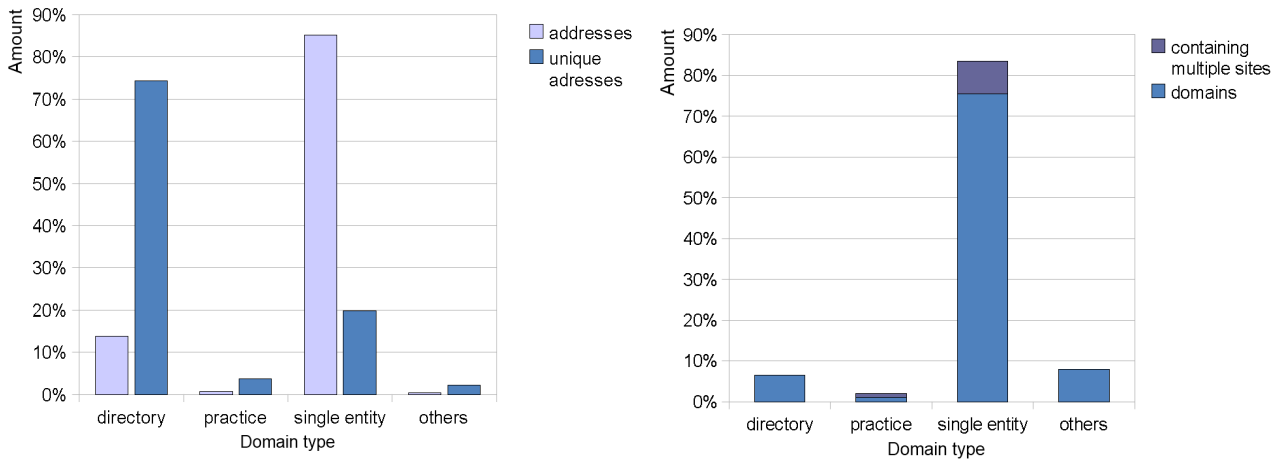

Figure 2: Distribution of addresses on domains: Count of addresses on domains by number of overall and unique addresses; Count of domains by type

entity [UCH03] by analyzing source frequency, domain name usage, name similarities etc. It can further be used to determine precedence of properties that are found on an official site and a third-party one.

After the entity merging and combination with directory data, entities are written back to the database. For a search of entities, an object-level ranking is employed on the features of the entities. The similarity score and potential markings and links are used to group entities where a clear resolution of conflicting data is not possible but which are still considered similar (cf. [JMCA07, CYC07]). This can often happen for multiple branches of a company, multiple owners of a practice or simply widely varying descriptions of actual same entities. We want to just pick one example of the data enrichment we undertake. As a way to improve the homepage classification, we use the number and distribution of overall addresses per domain. This allows to distinguish between directory-style pages, shared practices, and single entities (which might have multiple sites of branch offices) as seen in Fig. 2. A directory would contain lots of addresses which are mostly unique, a single entity domain might contain lots of addresses on all its pages, but they would mostly be the same, singling out the actual address.

Because the system is used in the backend, the requirement is that of very high recall and the support of manual quality assessment. Because of this, we rather err on the side of caution and use less aggressive matching and merging by adjusting the thresholds accordingly. To counteract the increase in near-similar entities, they are grouped together and also annotated with a reliability score to give the editors a quick overview of deviations or changes. This approach is sufficient to detect changes in known data that concerns the business properties of an entity. Additionally, changes of addresses are a very important fact. To capture this, an additional grouping dimension was introduced that adds entities with very similar names and other properties, but at different addresses, to the known entity from the provider's known entities.

2.4 Integration

The screenshots in Fig. 3 show the Web-based interface on the left and a detailed result entity. As the use case of the survey department is the integration of business data along a street-based workflow, the map component was deemed unnecessary. Instead, search by a selection of entity features was introduced. The most used feature is the search by an address or for a full street according to the workflow as well as searches for names to examine certain entities. However, in the final workflow, our system is just used as one input of many in specialized systems for manual inspection and semi-automatic data import. Therefore, an additional output was designed that delivered a table view that could be integrated into it and followed a certain format. A challenge here was to condense all the information about one entity into just one row with limited cells and also to express similarity of entities in a one-dimensional list, which was manageable because the similarity clusters are usually rather small at about 5 entries. In this format, the results are aggregated and assigned reliability scores to for an easy overview. We implemented a further ranking in this part to show the URL with the most reliability and also the most found properties for an entity to be accessible directly to the editors. All addresses are normalized to the spelling used in the gazetteer to make them disambiguous. Especially the grouping and reliability scores made

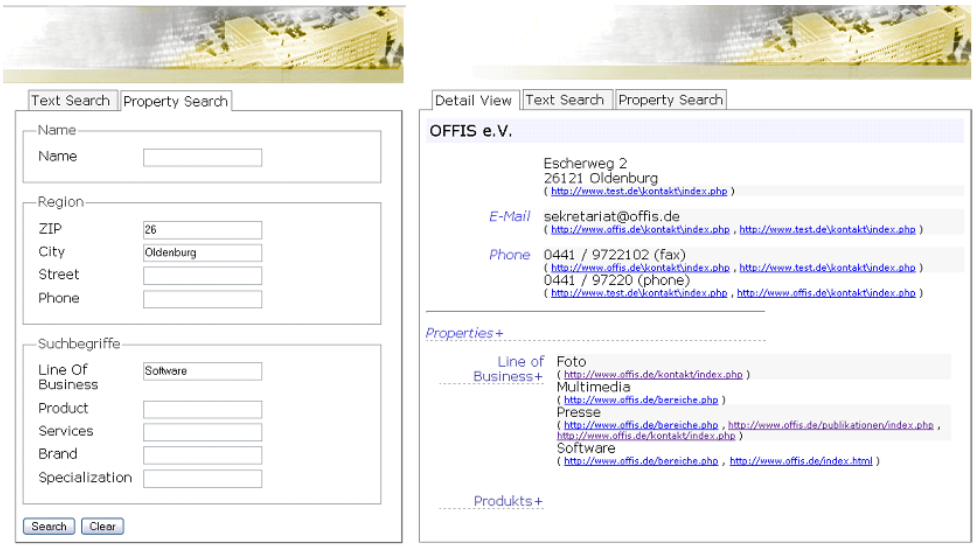

Figure 3: Web-based Entity Search Engine; Query Form and Result View

the integration a success because it made the high level of recall and thus a high number of results manageable and gave editors the ability to work through results in an ordered fashion so they were able to process all found entities.

3 Conclusion

On the backend, gazetteers for the covered cities feed a city-level geospatially focused resource discovery strategy. This allows to quickly create and set up a specialized local search engine for a confined region which can scan it reasonably fast and so quickly create regional search coverage. The process is driven by the integration of existing knowledge within the process. The gazetteer is used to generate the seeds and more importantly, to verify addresses. The existing client data is used to match newfound entities, assist in the merging and duplicate detection of entity data. The strength of the developed business entity search lies in the use of the complete text of the Web pages that allows extraction and subsequent search of properties and keywords usually not found within POI-databases or providers of database-backed local search such as queries for regional specialties, brand names, or other information. The main country-dependent components are the address-level geoparser and some extractors for, e.g., business registration numbers. This could be adapted to other countries with reasonable resources, a lager challenge would be the adaptation to countries with less precise address schemes such as lacking house numbers [AB08].

A 2008 study claims that 77% of businesses in Germany have their own Website [EUR08], but this means that more than 20% have no presence and actually, there are many smaller businesses that will not invest into a Web page. Contrary to [Him05], this means that some information is still missing. However, nowadays basic information can now also come from other places, such as online reviews or location-based servies' entries about purely offline shops. Thereby user-generated data can massively contribute to a better picture of local businesses, administrations, museums, or other points of interest for yellow pages.

The automatic data acquisition from business-oriented Web sites is used to improve quality and search time significantly in the survey department, since a much better and faster overview of companies is gained. Therefore, yellow pages are both a source and a beneficiary of local search technology.

Acknowledgements

Part of this work was done while at the OFFIS Institute for Information Technology, Oldenburg, Germany. This work extends a brief description of the project that was previously published in [Ahl11].

References

[AB07] Dirk Ahlers and Susanne Boll. Location-based Web search. In Arno Scharl and Klaus Tochterman, editors, *The Geospatial Web. How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society*. Springer, London, 2007.

[AB08] Dirk Ahlers and Susanne Boll. Retrieving Address-based Locations from the Web. In *GIR '08: Proceedings of the 5th International Workshop on Geographic Information Retrieval*, 2008.

[AB09] Dirk Ahlers and Susanne Boll. Adaptive Geospatially Focused Crawling. In *CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 445–454, 2009.

[Ahl11] Dirk Ahlers. *Geographically Focused Web Information Retrieval*, volume 18 of *Oldenburg Computer Science Series*. OlWIR, 2011. PhD Thesis.

[Ahl12] Dirk Ahlers. Local Web Search Examined. In Dirk Lewandowski, editor, *Web Search Engine Research*, volume 4 of *Library and Information Science*, pages 47–78. Emerald, 2012.

[AHSS04] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-Where: Geotagging Web Content. In *SIGIR '04*, 2004.

[BYCMR05] Ricardo Baeza-Yates, Carlos Castillo, Mauricio Marin, and Andrea Rodriguez. Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering. In *WWW '05*, pages 864–872, 2005.

[CKGS06] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18:1411–1428, 2006.

[CMH08] Michael J. Cafarella, Jayant Madhavan, and Alon Halevy. Web-Scale Extraction of Structured Data. *SIGMOD Rec.*, 37(4):55–61, 2008.

[CYC07] Tao Cheng, Xifeng Yan, and Kevin Chen-Chuan Chang. EntityRank: Searching Entities Directly and Holistically. In *VLDB '07*, pages 387–398. 2007.

[EUR08] EUROSTAT Press Office. Use of information and communication technologies – STAT/08/173. Technical report, EUROSTAT, 2008.

[GS08] Daniel Gomes and Mário J. Silva. The Viúva Negra crawler: an experience report. *Software: Practice and Experience*, 38(2):161–188, 2008.

[Him05] Marty Himmelstein. Local Search: The Internet Is the Yellow Pages. *IEEE Computer*, 38(2):26–34, 2005.

[JMCA07] Michael T. Jones, Brian McClendon, Amin P. Charaniya, and Michael Ashbridge. Entity Display Priority in a Distributed Geographic Information System, 2007. US Patent 20070143345.

[LKS08] Ryong Lee, Daisuke Kitayama, and Kazutoshi Sumiya. Web-based Evidence Excavation to Explore the Authenticity of Local Events. In *WICOW '08: Proceeding of the 2nd ACM Workshop on Information Credibility on the Web*, pages 63–66, 2008.

[MBS05] Alexander Markowetz, Thomas Brinkhoff, and Bernhard Seeger. Exploiting the Internet As a Geospatial Database. In *International Workshop on Next Generation Geospatial Information*, 2005.

[MGF09] Marcelo Mendoza, Hipolito Guerrero, and Julio Farias. Inquiro.CL: a New Search Engine in Chile. In *WWW '09 (WWW in Ibero-America track)*, 2009.

[UCH03] Trystan Upstill, Nick Craswell, and David Hawking. Query-Independent Evidence in Home Page Finding. *ACM Transactions on Information Systems*, 21(3):286–313, 2003.