Extraction of Business Information on the Web to Supply a Geolocated Search Service

Armel Fotsoh Tawofaing*, Christian Sallaberry†, Annig Le Parc - Lacayrelle†, Tanguy Moal‡
*LIUPPA - Cogniteev
Email: aftawofaing@univ-pau.fr
†LIUPPA, University of Pau, France
Email: christian.sallaberry@univ-pau.fr, annig.lacayrelle@univ-pau.fr
‡Cogniteev, Bordeaux, France
Email: tanguy@cogniteev.com

***Abstract*—Searching information about local businesses is not a trivial problem to address. Most of existing services are supplied with manually recorded data. Based on the observation that more and more businesses are referenced on the web, we propose a new approach, which consists to extract companies' targeted information (addresses, activities, jobs, products, emails, fax, phone numbers) from websites, to supply a local business search service. The information retrieval module combines thematic, spatial and full-text criteria.**

Keywords–Information Extraction; Geographic Information Retrieval

I. INTRODUCTION

Identifying specific information on the web according to a spatial localization is a topic increasingly explored. For example, a geolocated search service dedicated to emergency facilities is presented in [1]. Our research goal is to crawl the web and extract data in order to build specific geolocated entities, like businesses, events or persons.

This contribution presents the architecture of a service dedicated to local business search. As opposed to traditional search engines that relies on full document indexation, business local search services, mostly relies on companies descriptors provided by some specialized organisations. Far from adequate to build efficient systems, the basic descriptors must be supplemented by new ones. Therefore, we propose a service crawling the web, extracting information about companies (activity fields, practised jobs, commercialized or manufactured products, postal addresses, emails, phone and fax numbers) and storing them in indexes. The ultimate goal is to process a user need containing a thematic part (jobs, activities or products) and a spatial one, in order to query the indices and to get relevant results according to such different criteria.

The proposal relies on a model of business entity that is composed of two different parts. The first one is constituted of business basic data (official name, registration category, identifier, etc.) collected by crawling some specific administrative directories. The second part is composed of extended data extracted from companies' websites by using knowledge resource and pattern based extraction approaches. The retrieval system relies on the corresponding business entities and geolocated data.

The rest of this paper is organized as follow. Section II presents some related work; Section III describes the architecture of the proposed approach used to build our service; Section IV presents the implementation of a first version of a prototype and Section V concludes the paper and presents some prospects.

II. RELATED WORK

Some research works focus on the extraction of information related to businesses on the web. For example, Ahlers [2] develops a system which analyses web pages content in other to enrich a Yellow Pages [3] data provider. Analysed web pages are identified using Directory Mozilla (DMOZ) [4]. Extracted data here is addresses, phone numbers, emails, commercial and tax information of businesses. This data is used to consolidate and enrich the one contained in the Yellow Pages database. It is important to note that, in the French context in particular, the proportion of businesses registered in DMOZ remains very slight. Thus enriched data will also be limited.

Bootstrapping targeted data from Internet is a topic increasingly explored by several research works. Rae and al. [5] propose an approach for bootstrapping the web in order to identify Points Of Interest (POIs) websites. Their proposal uses Wikipedia data to list a POI information which is used to query Bing API as to retrieve the most relevant website corresponding to this POI.

Furthermore, many services dedicated to business information retrieval are available on the web. We organize these services into three main categories: (i) data providers like Factual [6] which collect and commercialize business data; (ii) directories like Yellow Pages, Google Maps [7] which contain a database of business information available online; (iii) social networks like Yelp [8] or Foursquare [9], which are services used in information sharing and reviews about businesses, places or events. Data supplying these services mostly come from manual recordings (users and employees), partners companies and open data. Thereby, if a company is not registered in the databases supplying those services or if its data is not updated frequently, it will lead us to have missing or out of date information.

Extraction of geographical information in web pages is a well-explored area of research. Some specific works focus on automatic extraction of addresses in unstructured web pages. An ontology-based approach for recognizing, extracting and geocoding Brazilian addresses in web pages is presented in [10]. A pattern-based approach [11] is used by Ahlers and Bool in [12] to present a technique of extraction and validation of

 ISBN: 978-1-61208-457-2

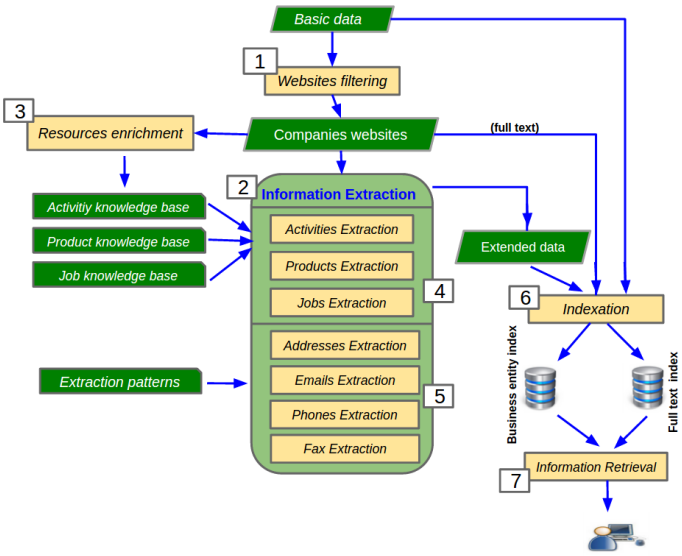

Figure 1. Processing chain

German addresses from web pages. City name or ZIP code are the two entry points of the extraction process. Moreover, [12] uses a street extraction process based on a gazetteer containing all the German street names.

Exploitation of knowledge resources is also a technique increasingly used for the extraction of thematic information in text, especially with significant progress in semantic web field these last years. Structuring and formalizing the knowledge of a specific domain in an ontology allows to annotate concepts [13] and semantic relations [14] in the text. Therefore, it is possible to perform semantic reasoning on the extracted information.

III. PROPOSITION

We propose an architecture of a local business search service, supplied with web data. Indeed, the contribution describes an approach that aims companies' websites identification on the web, and extraction of location and thematic information from these websites by combining some information extraction techniques. This proposal improves the existing services named in the state of the art, because it aims construction of low cost and up to date data. Differently from [2], we extract data in a web pages corpus constituted by filtering companies websites based on a heuristic. Besides of contact information and location data, we also extract activities, products and jobs in companies' web pages, using knowledge resources.
The process flow of our service is composed of four main steps (Figure 1).

A. Preprocessing

In this first step, one of our goals is to constitute the corpus of companies' websites in which we want to extract information. For this purpose, we have to bootstrap the web in order to filter those websites. We use a similar approach to the one described in [5] for our corpus constitution task. In fact, a directory containing license information of companies [15] is crawled to collect business basic data. These basic data, especially the name of the company and its localization, are used to query automatically Google and identify company website when it exists (Figure 1, Process 1). Business directories, social and professional networks are stored into a stop list in order to be filtered during the website search process. This helps to reduce not relevant results in the filtering process. Websites list identified during this step constitutes the input data to the extraction process.

Besides, we have constructed three core knowledge resources describing business activities, products and job positions respectively. These resources are based on hierarchical organisations defined by the French National Statistics Institute called INSEE [16] for the first two ones and by the French organisation for work, Pole Emploi [17] for the third one. We transformed all these resources in order to represent them as OWL ontologies. Furthermore, the Latent Dirichlet Allocation (LDA) clustering algorithm [18] is applied to companies' websites of each activity group, in order to extract the corresponding vocabulary. Indeed, this learning algorithm enriches our core knowledge resources (Figure 1, Process 3).

B. Information extraction in websites

One of the most difficult challenges in web pages analysis is that information on Internet is mostly unstructured. Indeed, only few websites use metadata or microformats to publish structured information. In our proposal, we want to analyse companies' websites and extract thematic and spatial information in order to fill in business extended data.

1) Thematic information extraction: Extraction of activities, jobs and products relies on an ontology-based approach. The three knowledge resources built and enriched in the preprocessing step, are used to annotate automatically the website corpus (Figure 1, process 4). Each term or phrase in pages content which corresponds to a resource category is tagged with the corresponding identifier.

Emails, phone and fax numbers are extracted by using a pattern-based approach. We wrote extraction rules by observing patterns used for the writing of the targeted information in a sample of French companies' websites. Phone numbers follow a specific pattern, depending if it has a country telephone code or not. Our proposal makes a distinction between mobile and landline phone numbers based on French standards. Fax number are landline phone numbers introduced by special keywords ("Télécopie", "Télécopieur", "Fax", etc.). An email is a phrase which follows this pattern (Table I corresponds to the legend of extraction patterns) :

$$\textbf{email} \rightarrow Login \quad (\text{"@"} \mid \text{"}(at)\text{"}) \quad Domain\_Name \quad (\text{"."} \mid \text{"}(dot)\text{"}) \quad Domain\_Extension$$

The entry point of the extractor is the identification of a domain extension ($Domain\_Extension$) and "@" or "(at)" symbol. We use a gazetteer of domain extensions for this purpose.

2) Spatial information extraction: In the literature, models are proposed to represent entities like addresses. Schema.org has a module dedicated to address modelling. In our business model, location representation is based on the one defined by the French governmental data lab named Etalab [19]. Our goal is to extract addresses in web pages up to the street number granularity level according to the Etalab addresses model.

 ISBN: 978-1-61208-457-2

TABLE I. LEGEND OF EXTRACTION PATTERNS

A ?	A is optional
A | B	A or B
A B	A and B
A(n)	A is repeated n times
[i-j]	element in set i, ..., j with $i < j$

We also want to extract information like address supplements (building name, floor number, etc.), postal boxes numbers and letter numbers (it is a number used by postal services to facilitate postal mail transmission, e.g., "CEDEX 07").

Several difficulties are related to addresses extraction from text in general and French ones in particular. In fact, there are many address formats used and standards about the order of the various components of an address do not exist (a council name is not always following a ZIP code). Besides, just a few web publishers use address keyword introducer or specific tagging to facilitate address automatic identification. Furthermore, in the French context, there is no a complete and available gazetteer containing all street names. Thereby, the approach proposed by [12] has to be adapted in our study case.

We propose an approach to automatically extract French addresses in web pages. The observation of a sample of 160 companies' websites allowed us to identify some frequent patterns of addresses formats in French companies' websites. In this sample, the ZIP code was always present in the identified addresses. Table II describes the different components used in the expression of a French address. Each one is extracted using gazetteers or specific rules. For example, ZIP Code is extracted using the following rules:

$$\mathbf{ZC} \rightarrow \ "F-" \quad [0-9](5) \qquad (1)$$
$$\mathbf{ZC} \rightarrow \ [0-9](5) \qquad (2)$$
$$\mathbf{ZC} \rightarrow \ [0-9](2) \quad [0-9](3) \qquad (3)$$

Priorities are associated to each of these rules. The rule (1) has the highest invocation priority, rule (2) afterwards. More complex rules are used to extract fields like street name (SNa) and address supplement(AS).

The address extraction patterns observed have been summarized into three main rules. They are described below:

Extraction rule 1

$$\mathbf{Address} \rightarrow \ AS? \quad ((PB \quad SC) \mid (SC \quad PB) \mid PB \mid SC)? \quad SNu? \quad SNa \quad AS? \quad ((PB \quad SC) \mid (SC \quad PB) \mid PB \mid SC)? \quad ((ZC\ C) \mid (C\ ZC)) \quad LN? \quad D? \quad Co?$$

Extraction rule 2

$$\mathbf{Address} \rightarrow \ AS? \quad ((PB \quad SC) \mid (SC \quad PB) \mid PB \mid SC)? \quad ((ZC\ C) \mid (C\ ZC)) \quad LN? \quad D? \quad Co?$$

Extraction rule 3

TABLE II. ADDRESS COMPONENTS

Field names	Symbols	Examples
Address Supplement	AS	Résidence Rigaud
Postal Box	PB	BP 1167
Special Course	SC	CS 2587
Street Number	SNu	10 ter
Street Name	SNa	Avenue de l'université
ZIP Code	ZC	64000
City Name	C	Pau
Letter Number	LN	CEDEX 01
Department	D	Pyrénées-Atlantiques
Country	Co	France
Street Introduce	SI	Avenue

$$\mathbf{Address} \rightarrow \ AS? \quad ((PB \quad SC) \mid (SC \quad PB) \mid PB \mid SC)? \quad SNu? \quad SNa \quad AS? \quad ((PB \quad SC) \mid (SC \quad PB) \mid PB \mid SC)? \quad ZC \quad LN? \quad D? \quad Co?$$

In the first pattern, the street name, ZIP code and city name are required (e.g., "10 Rue du Maréchal Foch 49000 Angers"). This pattern represents about 75% of the addresses of our dataset. In the second pattern, only the ZIP code and city name are required, street name must be omitted. (e.g., "Résidence Rigaud 33350 Mouliets-et-Villemartin"). This pattern represents about 14% of the observed addresses. In the third pattern, street name and ZIP code are required and the city name must be omitted ((e.g., "10 rue du Maréchal Foch F-33500"). This last case represents less than 4% of the detected addresses. Others components like the address supplement and postal box might complete the extracted addresses.

Algorithm 1 details the conditions of the different address extraction rules triggering. The entry point of the algorithm is the ZIP Code which is identified by using rules defined previously. For each sentence in which a potential ZIP Code is identified, if a council name and a street introducer are detected, pattern 1 is triggered. Otherwise, if a country name only is detected, pattern 2 is triggered. Finally if a street introducer only is detected, pattern 3 is triggered. In all the other cases, there is no address in the sentence. Let us note that SI detection is identified using a gazetteer.

Algorithm 1 Address Extraction Algorithm

```
for each sentence do
  if (ZC & C & SI) then
    trigger extraction rule 1
  else
    if (ZC & C) then
      trigger extraction rule 2
    else
      if (ZC & SI) then
        trigger extraction rule 3
      else
        No Address
      end if
    end if
  end if
end for
```

3) Full-text extraction: The content of each web page of the corpus is processed by eliminating all the HTML tags in page content as well as metadata. JavaScript scripts and CSS code are also removed. Only the full-text is kept.

C. Indexation

For each company, basic and extended data are merged to build a complete business entity according to the proposed model. Every extracted address is geocoded using Etalab tool. The business entities so constructed are stored in an index. In parallel, the full-text corresponding to each web page is stored in another index (Figure 1, process 6).

D. Information Retrieval

The indices built in the previous stage are used to answer user needs (Figure 1, process 7). The information retrieval process analyses each query to handle separately or accordingly its spatial and thematic parts. Furthermore, the retrieval process integrates the querying of the web pages whole content. This retrieval stage combines spatial, thematic and full-text querying criteria.

IV. Implementation

A first prototype of our approach has been implemented, dealing with companies of the South West region of France. We worked on 22,000 companies websites representing 212 business activity fields defined by INSEE. The related corpus is constituted by crawling all these websites with Apache Nutch [20] framework. This generates a corpus of 550,000 web pages. Both of the pattern-based and the ontology-based extraction approaches, described in the extraction process, are performed on this corpus using GATE [21] platform. We connected GATE with Apache HADOOP [22] framework in order to process the large volume of web pages. The use of HADOOP Map Reduce framework with two machines working in parallel, has reduced by more than 50% the time of web pages annotation. The information extracted in these web pages has been indexed using Elasticsearch [23]. The business entity and full-text indices have a total size of 3 GB.

”Oak beams in south of Bordeaux” is an example of user need. The system has to return business entities of the ”carpentry work” activity located in the south of Bordeaux. This information need is processed according to the Elasticsearch syntax and submitted to our prototype. The execution of the preview query retrieves a list of relevant business entities (the first one is http://www.belles-toitures-girondines.com).

V. Conclusion

Our service uses a modular structure. It combines several research areas and techniques, which become complementary in the construction of business entities. This solution leverages learning techniques to enrich knowledge resources. It also performs information extraction (spatial and thematic) using a pattern based approach and knowledge resources. Moreover, our service relies on a new model of business entities, combining administrative data and those extracted from websites. Finally, the service addresses spatial, thematic and full-text information retrieval.

Future work will focus on the evaluation of website filtering and information extraction processes. A definition of an efficient information retrieval model to combine such criteria and support natural language queries will also be carried out in future research. Moreover, an evaluation of the entire architecture with a representative set of queries will also be performed in future work.

References

[1] W. Li, M. F. Goodchild, R. L. Church, and B. Zhou, “Geospatial data mining on the web: Discovering locations of emergency service facilities,” 2012, pp. 552–563.

[2] D. Ahlers, “Business entity retrieval and data provision for yellow pages by local search,” in IRPS Workshop@ ECIR2013, 2013.

[3] “http://www.pagesjaunes.fr,” 2015.10.29.

[4] “http://www.dmoz.org,” 2015.10.29.

[5] A. Rae, V. Murdock, A. Popescu, and H. Bouchard, “Mining the web for points of interest,” in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’12. New York, NY, USA: ACM, 2012, pp. 711–720. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348379

[6] “http://www.factual.com,” 2015.10.29.

[7] “http://www.google.fr/maps,” 2015.10.29.

[8] “http://www.yelp.fr,” 2015.10.29.

[9] “http://fr.foursquare.com,” 2015.10.29.

[10] 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 Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, ser. GIR ’07. New York, NY, USA: ACM, 2007, pp. 31–36. [Online]. Available: http://doi.acm.org/10.1145/1316948.1316957

[11] S. Blohm, Large-scale pattern-based information extraction from the world wide web. KIT Scientific Publishing, 2011.

[12] D. Ahlers and S. Boll, “Retrieving address-based locations from the web,” in Proceedings of the 2Nd International Workshop on Geographic Information Retrieval, ser. GIR ’08. New York, NY, USA: ACM, 2008, pp. 27–34. [Online]. Available: http://doi.acm.org/10.1145/1460007.1460015

[13] S. Nešić, F. Crestani, M. Jazayeri, and D. Gašević, “Concept-based semantic annotation, indexing and retrieval of office-like document units.” CID, 2010.

[14] A. Royer, C. Sallaberry, A. Le Parc-Lacayrelle, and M.-N. Bessagnet, “Extraction automatique de relations sémantiques définies dans une ontologie,” in Actes RISE 2015, 2015, pp. 30–42.

[15] “http://www.societe.com,” 2015.10.29.

[16] “http://www.insee.fr/fr,” 2015.10.29.

[17] “http://www.pole-emploi.fr,” 2015.10.29.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, 2003, pp. 993–1022.

[19] “http://www.etalab.gouv.fr,” 2015.10.29.

[20] “http://nutch.apache.org,” 2015.10.29.

[21] “http://gate.ac.uk,” 2015.10.29.

[22] “http://hadoop.apache.org,” 2015.10.29.

[23] “http://www.elastic.co,” 2015.10.29.

 ISBN: 978-1-61208-457-2