A process for business entities extraction on the web

Armel Fotsoh, Christian Sallaberry, Annig Le Parc-Lacayrelle, Tanguy Moal

A process for business entities extraction on the web

Armel Fotsoh Tawofaing, University of Pau - COGNITEEV
Christian Sallaberry, University of Pau
Annig Le Parc - Lacayrelle, University of Pau
Tanguy Moal, COGNITEEV

Abstract

Searching information about local businesses is a difficult task. Most of existing services are supplied with manually recorded data, however, an increasing number of companies are referenced on Internet and release information on their websites. In addition, data collected from companies is made available by the administration as open data. Therefore, we propose a process to extract companies information such as addresses, activities, jobs, products, emails, fax and phone numbers from websites in order to offer a business search service with low cost constructed and updated data. This process relies on the use of knowledge-based and pattern-based extraction approaches. The proposal is composed of two main modules : the first one relies on a heuristic that uses companies registration data to bootstrap the web in order to filter their official corporate websites; the second module, on the other hand, analyses these websites to extract targeted data in order to map it on a dedicated knowledge graph, made of several indexes.

I. Introduction

Building local-based services appears to be a trend these days as it attracts an increasing interest. In fact, local-based services help users to find local services offers in their neighbourhood or around a specific spatial localization. In that sense, some services are dedicated to businesses and services geolocated search, such as Google My Business[1] and "Pages Jaunes"[2]. These services are however mostly supplied with manually recorded data, meaning there is no real time update and some data can be missed. For instance, a search for zinc-work companies in the French council of "Pujols Sur Ciron" on Google Maps does not lead to any result in the targeted council; nor does it have any results in "Pages Jaunes". As a matter of fact, "Ets JEROME BELLIN" is a zinc-work company indeed located in the targeted council of "Pujols Sur Ciron" and has its website[3] up-to-date with the current business information and location. This highlights the limitation of existing services dedicated to local businesses.

Some companies' information, such as their registration number, is freely provided by local or national institutions. In addition, an increasing number of companies are referenced on the web. According to "La Tribune" blog [4], in 2013, about 64% of French companies had an official website with the company information. In Europe, the presence rate is actually higher. Based on these observations, we propose an approach to address extraction of companies data on the web such as activity fields, practised jobs, commercialized or manufactured products, postal addresses, emails, phone or fax numbers.

The proposal relies on a model of business entity with two main parts; the first part is made of company registration data (official name, business category, identifier, managers) contained in dedicated open data websites. The second part is constituted of extended data extracted from companies' websites. Address extraction is one of the most difficult task in the process, due to the fact that micro data or specific tags are used just by few websites for publishing addresses. Besides, people use many different ways to write addresses.

The paper is organised as follow. Section II presents some related work; Section III describes our proposal; Section IV presents the implementation of our approach; Section V is about evaluation of the website filtering module and the information extraction one and Section V is the paper conclusion as well as some prospects.

II. Related Work

Few researchs focus on spatial named entities recognition in text documents. Vaid et al. [12] presents, an approach for the extraction and disambiguation of spatial named entities on the web. Their approach also allows the extraction of relative spatial entities introduced by topological expressions like "near of", "in the south of", etc. More specific works focus on the automatic extraction of addresses in textual documents. There are three mains techniques: (i) the first one is ontology-based. It is used by Borges et al. [4] for the recognition, extraction and geocoding of Brazilian addresses in web pages; (ii) the second one relies on machine learning techniques for the extraction of urban addresses. A learning algorithm (CRF: Conditional Random Field) is used in [5] for example to extract addresses in web pages; (iii) the last one is a pattern-based technique, it is used by Ahlers and Boll [2] for the extraction and the validation of German addresses from web pages and it relies on the use of many Gazetteers, one of which contains all German street names.

The use of knowledge resources in thematic text annotation is explored by several researchs. In fact, recent development in semantic web allows the structuring of the knowledge of a specific field in specialized resources like ontologies, thesauri, etc. The contribution described in [11], presents an approach for

[1] http://www.google.com/business/
[2] http://www.pagesjaunes.fr
[3] http://www.charpente-bellin.fr/
[4] http://www.latribune.fr/blogs/strategie-marketing-en-1min30/20130515trib000764662/64-des-entreprises-francaises-disposent-d-un-site-internet.html

textual documents annotation using a Football domain ontology. This formalization of a field knowledge in a structured organisation like ontologies allows the annotation of concepts [6] and/or relations between them [10] in text. Sometimes, knowledge resources do no always contain enough vocabulary for text annotation. Therefore, several techniques have been designed to enrich these resources. Parekh et al. [7] proposes an approach based on text mining to enrich, in a semi-automatic way, an ontology describing the medical field. They use medical textual documents and glossaries. The approach presented in [7] uses matrix decomposition in singular values to build clusters of terms from a selected input corpus. An expert of the domain validates selected terms for each cluster.

Discovering of web pages that describe specific entities on the web, is a field of research in full development. Rae and Murdock [8] propose a technique for bootstrapping points of interest (POIs) data from the web. In fact, their approach uses Wikipedia data, to collect a POI information and use it to find the most relevant corresponding website. Complementary information about the POI is then extracted from the selected website.

Many services dedicated to business information retrieval are available on the web. Data providers like Factual[5], collect and commercialize business information. Directories like "Pages Jaunes" or Google My Business are used to query business databases while social networks like Yelp[6] or Foursquare[7] are used in information sharing and reviews about businesses. All these services are supplied mostly by manually recorded data (end-users and employees), partner companies data or open data. Thus, data is not always up to date and some can be missed. Therefore, some works focus on business information extraction on the web. For instance, Alhers [1] proposed a system which analyses web pages content in order to to consolidate and enrich those contained in the Yellow Pages database. The analysed corpus is constituted of websites of companies referenced within Mozilla Directory (DMOZ[8]). Extracted data here are addresses, using the approach details in [2], phone numbers, emails, commercial and tax information of businesses. Note that, the proportion of existing companies registered in DMOZ is very low, therefore the extracted data remains poor and the wealth of companies information available on the web is only partially exploited.

III. Proposition

The process flow of our service is composed of several main steps:

[5]http://www.factual.com,
[6]http://www.yelp.fr/
[7]http://fr.foursquare.com
[8]http://www.dmoz.org

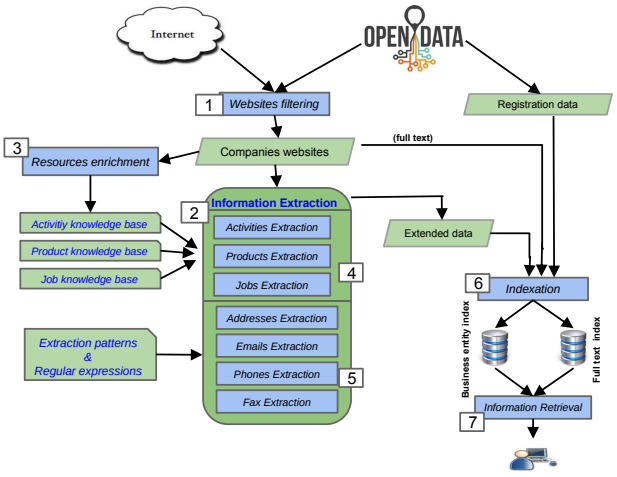

Figure 1: *Processing chain*

- After collecting companies' registration information coming from open data platform, a first step (Figure 1.1) filters theirs websites on Internet. Our approach is different from the one used by Ahlers [1] in the way that, websites constituting extraction corpus are filtered directly from the web, using an heuristic similar to the one used in [8]. Hence, the number of websites used for the information extraction process is more significant. Besides contact information and addresses, we also extract companies activities fields, products and practised jobs from the websites using knowledge resources.

- An other step (Figure 1.3) based on the approach detailed in [7], is used for knowledge resources enrichment.

- Target information is automatically extracted from those websites' contents using pattern-based and ontology-based approaches(Figure 1.2).

- Business entities are constructed from registration and extracted data. These constructed entities are then stored in the knowledge graph (Figure 1.6). Our approach is designed for French companies, but it can also be adapted for others countries.

Next sub-sections detail data and resources preprocessing, website filtering, information extraction and business knowledge graph storage.

I. Open data Leverage

Open data seems to be a trend for data publishing. Access to data is more and more generalized with the real revolution brought by the semantic web. Even public administration are not left out. They publish a huge volume of information in open access on the web. "Trade and Companies Register" for example collect, publish and maintain French' companies

registration information on an open data platform[9]. That information is structured according to the standard model of the SIRENE repository defined by the French National Statistics Institute INSEE[10]. This registration data is the entry point of our website filtering process. The SIRENE model is also use to represent registration data in a business entity.

INSEE has also defined two hierarchies for organizing companies' activity fields and manufactured or commercialized products. This is respectively the "French Activities Nomenclature" (NAF) and the "French Classification of Products" (CPF). The French organisation for work Pole Emploi[11] has defined an hierarchy, "Trade and Jobs Operational Register" (ROME) for organizing the different socio-professional categories. These hierarchies are published on their related institution websites in open access. They are used in our proposal for information extraction purpose.

Besides, the French governmental laboratory of open data Etalab[12], defines a consensus model of postal addresses representation. Addresses, according to the defined model, are geocoded and published in open access on a dedicated platform[13]. Our proposal uses this open data platform to geocode detected addresses in websites.

Open data is used in the process flow stage for filtering companies' websites, building knowledge resources for information extraction task and defining the structure of data. Table 1 presents a simplified model of entities in the business knowledge graph. For each property, we also provide which data source has been used.

Table 1: *SIMPLIFIED BUSINESS ENTITY MODEL*

Business Entity Properties	Representation Models	Data Sources
Registration data	SIRENE (INSEE)	datainfogreffe.com
Coordinates		
website	-	web
Address	Address (Etalab)	company website
Phone Numbers, emails, fax	-	company website
extended data		
Jobs	ROME (Pole Emploi)	company website
Activities	NAF (INSEE)	company website
Products	CPF (INSEE)	company website

II. Website filtering

Registration data coming from datainfogreffe.fr is used to query the web in order to identify companies' websites. In fact, datainfogreffe.fr is a platform supplied and maintained by the public administration, thereby, companies name and main activity data we need are reliable. An heuristic based on this registration data is defined to filter companies' websites. Figure 2 details the heuristic. Indeed, several combinations of registration data are tested in order to identify websites from a sample of 300 French companies. A combination of some properties was sufficient to identify a company website when it exists. *Commercial name* and *the council name* is the most effective combination. If there is no commercial name, replacing it with the $official\ name$ gives quite positive answer. A company necessarily has an official name (e.g. "ETABLISSEMENTS PIERRE DURIEZ"), it might also have a commercial name (e.g. "DURIEZ AGENCEMENT"), but it is not mandatory. The algorithm 1 presents our website filtering heuristic. A blacklisted websites list, containing business directories, council websites, people directories, social networks and administration websites is constituted and used in the heuristic to reduce errors.

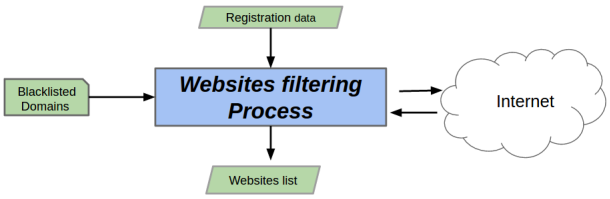

Figure 2: *Website Filtering*

Algorithm 1 Website Filtering

Require: blacklisted websites
  for each company of the trade register **do**
    $website \leftarrow Null$
    $website\_found \leftarrow$ False
    if ($commercial\_name$ exists) **then**
      $step \leftarrow 1$
    else
      $step \leftarrow 2$
    end if
    while ($step < 3$) and ($not(website\_found)$) **do**
      if ($step = 1$) **then**
        $query \leftarrow commercial\_name + council\_name$
      else
        $query = official\_name + council\_name$
      end if
      execute *query* and retrieve the three most relevant websites
      if all results in blacklisted websites **then**
        $step \leftarrow step + 1$
      else
        $website\_found \leftarrow$ True
        $website \leftarrow$ first not blacklisted website
      end if
    end while
  end for

[9] http://datainfogreffe.fr
[10] http://www.insee.fr/fr
[11] http://www.pole-emploi.fr
[12] http://www.etalab.gouv.fr
[13] http://adresse.data.gouv.fr/

III. Resources enrichment

For the automatic extraction of activities, jobs and products, we use three knowledge resources. These resources are built by transforming the NAF, ROME, CPF hierarchies as OWL ontologies. The three core hierarchies choice is motivated by their organizational level which is fairly atomic. Each ontology used in our proposal, is a graph of concepts (activity, jobs or products) with hierarchical relations. A set of vocabulary labels is used to describe each concept.

The activity ontology built from the NAF hierarchy is poor in vocabulary labels. Therefore, we decided to use the website corpus to enrich it. We propose a semi-automatic text-mining technique similar to the one describes in [7]. Indeed, registration data of companies contains information about its main activity field. The idea is to use a learning algorithm to gather phrases in companies' websites of a same activity field. An expert will validate the selected phrases and they will be added to the core resource. Given the volume of companies' websites available for each activity field which is quite low, the Latent semantic Allocation (LSA) algorithm does not give good results for such a vocabulary gathering task. However, Latent Dirichlet Allocation (LDA) [3], allows us to obtain more acceptable results. Therefore, this last algorithm is used for our clustering purpose. It is also important to mention that phrases can be uni-gram, 2-gram, 3-gram and 4-grams.

For a given activity field, a corpus of all associated web pages is constituted. In this corpus, there are phrases related to the activity field vocabulary as well as others latent vocabularies. The goal is to gather web pages phrases into vocabulary clusters. The clustering function takes several parameters :

- **n**: which is the size of phrases in each vocabulary cluster (1 to 4).
- **V**: which is the vocabulary of the corpus, it is constituted of all the distinct n-grams.
- **M**: which is the matrix representing relations between each phrase of the vocabulary and corpus pages. Columns of M are constituted of V elements and lines of web pages. M_{ij} is the weight of the j phrase in the i web page. In our proposal, the weighting is based on an adaptation of the TF-IDF [9].
- **nb_topics**: which is the number of phrases' clusters we want to extract from the websites corpus.
- **nb_top_words**: which is the number of top phrases in each cluster.

As output, we have clusters of phrases corresponding to topics identified in corpus. An expert selected relevant phrases from the cluster corresponding to the activity field description. These phrases are then added to the associated class in ontology.

IV. Information extraction

Each web page is analysed to annotate targeted information (Figure 1.2). The annotations are extracted from web pages to fill in extended data properties (1) .

Thematic information extraction Extraction of activities, jobs and products relies on an ontology-based approach. We use the three knowledge resources built and enriched in III.III to automatically annotate web pages. These ontologies are processed as follow: vocabulary labels are lemmatized and case normalized for the semantic annotation task. Besides, each web page content is also tokenized, lemmatized and case normalized. Thereby, web page phrases matching an ontology class label are tagged with the corresponding class identifier.

Regular expressions are used to annotate emails, phone and fax numbers in web pages. Observation of this information in a sample of French companies websites allowed us to write the extraction patterns.

Table 2: *ADDRESS COMPONENTS*

Field names	Symbols	Examples
Address Supplement	AS	Résidence Rigaud
Postal Box	PB	BP 1167
Special Course Number	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

Address information extraction The System described in [2], uses a Gazetteer of all German streets to extract addresses in web pages. Thus, the precision of the system is pretty good. In the French context, a Gazetteer with an exhaustive list of street names is not available. Therefore, we propose a new pattern-based approach for address extraction from web pages. Observation of a sample of 160 French companies websites allowed us to define a set of address patterns. The Zip Code is almost always presented, thereby, it is the entrance point of the address extraction process. Table 2 details the different components of a French address in its full form. Each component is identified by dedicated patterns. When an address contains at least ZIP code, council name and street information, the extraction pattern is the one bellow. Other information such as Address Supplement, Postal Box or Special Course Number might be present before or after the street name.

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

Legend

$$A? : A \text{ can be omitted}$$
$$A \mid B : A \text{ or } B$$
$$A\quad B : A \text{ and } B$$

V. Indexation

Annotations are extracted from web pages to construct corresponding extended data. Geocoding of extracted addresses is performed before merging extended data with registration one to construct business entities according to the model defined in III.I. All the business entities so obtained are added to an index representing the business knowledge graph while textual content of webpages is stored in another full text index for the search module.

IV. Process flow implementation

Our prototype deals with the South West region of France (Aquitaine) where about 254,000 companies are registered. We chose 212 INSEE activity fields corresponding to about 115,000 companies: 22,000 of them have a website according to the website filtering heuristic. We use Google Search API[14] for executing search queries. Corpus is constituted by crawling all these websites with Apache Nutch Framework[15] (we crawl only the home page and the first level pages for each website). This produces a corpus of about 550,000 web pages.

We use the Python library lda[16], which implements the clustering function described in III.III. This process allows the enrichment of the activity ontology with new class phrases. Corpus annotation is performed using GATE[17] Platform. In order to deal with French language, we integrate TreeTagger in GATE. We also connect GATE to Apache HADOOP[18] framework in order to deal with the large volume of web pages. Thus, the annotation process is scalable and its implementation does not depend on data size. The knowledge graph of business entity is indexed using Elasticsearch[19].

For the corpus of 550,000 web pages, our annotation process identified about 30,000 addresses, 44,000 activity phrases, 12,500 product phrases and 28,000 job phrases. The business and the full text indexes have a total size of 3 GB.

[14] http://developers.google.com/custom-search/json-api/v1/overview
[15] http://nutch.apache.org
[16] http://pypi.python.org/pypi/lda
[17] http://gate.ac.uk
[18] http://hadoop.apache.org
[19] http://www.elastic.co

V. Evaluation

This section details the evaluation of some process flow steps such as website filtering and information extraction.

I. Website Filtering Evaluation

A set of 500 entities are randomly selected from the business knowledge graph for website filtering evaluation. A manual verification is then carried out to analyse each company website in a browser to validate if it is relevant or not. About 30% of the retrieved websites were not relevant.

These results can be explained by the fact that some not relevant domain names are not registered in our blacklisted websites list. Moreover, many companies, especially artisans, have a commercial or official name which coincide with the founder one. Therefore, the website filtering process brings back no relevant websites, such as magazines, newspapers, universities, institutions etc. as long as the founders first name or last name appears in these websites.

II. Information Extraction Evaluation

We evaluate both the address and the activity fields extraction processes. There is no evaluation campaign available dedicated for this type of entities (address, activity), therefore, we have defined one. Our campaign follows the classic TREC process like the one described in [13]. It contains:

- Categories: each one corresponds to the entity type we want to tag in the corpus (Address or Activity).
- Pages: it is constituted of the information needs in the TREC approach. In our case, each page contains one or more addresses (activities respectively).
- Query relevance judgements (qrels): this is the mapping between each page and the relevant extracted entities. In fact, an expert has annotated in each page all the relevant addresses (and activities respectively).

The evaluation corpus is submitted to the annotator system. Extracted entities and qrels are used to calculate each page precision and recall. Means of these metrics, on all the evaluation corpus, give respectively the global precision and the global recall. The harmonic mean of these global metrics corresponds to the F-measure.

II.1 Address Extraction Evaluation

The evaluation corpus is constituted of 240 web pages containing at least one address. 309 addresses are annotated by the expert, while 286 are selected by our annotator. Among those selected by the annotator, 251 are judged relevant and 13 have been partially selected (some elements are missed). Table 3 presents summary results of address evaluation process.

Table 3: *ADDRESS EVALUATION SUMMARY*

Relevant	Selected	Precision	Recall	F-Measure
309	286	0.80	0.81	0.80

Several selected addresses are not relevant. Many reasons explain this observation: council name is not in the council Gazetteer or it is misspelled, this raises mostly with composed council name ("Bon-Encontre"); the street name contains a council name and the correct council name is not in the council Gazetteer ("Rte de Martillac 40200 Bon-Encontre").

However, some relevant addresses are not selected, due to the fact that they are not covered by our extraction patterns. It is the case when the Zip Code is omitted ("10 Place de la Bourse Bordeaux") and sometimes the council name too ("10 Place de la Bourse").

II.2 Activity Extraction Evaluation

We have constituted a corpus of 100 company websites' home pages. We deal only with activities related to roof (carpentry, covering, etc.). The system annotates 1119 activity phrases while the expert annotates 1131. Only 631 among the 1119 activity phrases selected by our annotator are correct. Table 4 summaries results of the activity evaluation process.

Table 4: *ACTIVITY EVALUATION SUMMARY*

Relevant	Selected	Precision	Recall	F-Measure
1131	1119	0.45	0.52	0.49

About half of the pertinent activities are not annotated. This is mostly due to the poorness of the activity ontology is spite of its enrichment. Moreover, many of the activity labels do not represent all the possible variations of a phrase. Indeed, the expert can annotate a label which is only partially recorded in the ontology. He can identify for example the expression "traditional wood frame", while only "wood traditional frame" is recorded as a label in the ontology. In addition, some errors may be due to misspellings in web pages or even phrases lemmatization problems during the annotation process.

VI. Conclusion

We propose a modular process dedicated to the extraction of business thematic and location information on the web. This process relies on several research areas which are combined in a complementary way to construct business entities. These entities are stored in a business knowledge graph. The proposed approach aims the extraction of addresses in websites in spite of the different expression formats, with quite good effectiveness results. Despite the poor results obtained, the enrichment of knowledge resources with learning techniques, contributes to improve thematic information extraction in websites. Our model of business entities is based on the one used by the administration to publish open data combined with those defined by specialized organisations.

Filtering websites, which is a key stage in the process, has to be improved. Therefore, future work will explore how to add a classifier in the filtering step in order to automatically detect if a website is a directory or a company one. Other future work will also focus on the semantic disambiguation of activities, products and jobs detected in a company' website. This will help to keep only in the knowledge graph relevant information for each entity. Finally, we will focus on the definition of information retrieval models combining spatial, thematic and full text criteria for business local-based search.

References

[1] Dirk Ahlers. Business entity retrieval and data provision for yellow pages by local search. In *IRPS Workshop@ ECIR2013*, 2013.

[2] Dirk Ahlers and Susanne Boll. Retrieving address-based locations from the web. In *Proceedings of the 2Nd International Workshop on Geographic Information Retrieval*, GIR '08, pages 27–34, New York, NY, USA, 2008. ACM.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] Karla A. V. Borges, Alberto H. F. Laender, Claudia B. Medeiros, and Clodoveu A. Davis, Jr. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pages 31–36, New York, NY, USA, 2007. ACM.

[5] Berenike Loos and Chris Biemann. supporting web-based address extraction with unsupervised tagging. In *Data Analysis, Machine Learning and Applications 2008*, pages 577–584, 2008.

[6] Saša Nešić, Fabio Crestani, Mehdi Jazayeri, and Dragan Gašević. Concept-based semantic annotation, indexing and retrieval of office-like document units. CID, 2010.

[7] Viral Parekh, Jack Gwo, and Tim Finim. Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *International Conference of Information and Knowledge Engineering*, 2004.

[8] Adam Rae, Vanessa Murdock, Adrian Popescu, and Hugues Bouchard. Mining the web for points of interest. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 711–720, New York, NY, USA, 2012. ACM.

[9] François Rousseau and Michalis Vazirgiannis. Graph-of-word and tw-idf: New approach to ad hoc ir. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 59–68, New York, NY, USA, 2013. ACM.

[10] Albert Royer, Christian Sallaberry, Annig Le Parc-Lacayrelle, and Marie-Noëlle Bessagnet. Extraction automatique de relations sémantiques définies dans une ontologie. In *Actes RISE 2015*, pages 30–42, 2015.

[11] Kara Soner, Alan Ozgur, Sabuncu Orkunt, Akpinar Samet, Cicekli Nihan K., and Alpaslan Ferda N. An ontology-based retrieval system using semantic indexing. volume 37, pages 294–305, Oxford, UK, UK, June 2012.

[12] Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual indexing for geographical search on the web. In *Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, pages 218–235, Berlin, Heidelberg, 2005.

[13] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press, 2005.