Automating the Shaping of Metadata Extracted from a Company Website with Open Source Tools

DR IR ROBERT VISEUR CETIC
Rue des Frères Wright, 29/3 6041 Charleroi, Belgium
UMONS Faculty of Engineering
Rue de Houdain, 97000 Mons,
Belgium

***Abstract* — As part of a market analysis process, the objective was to automate the task of identifying the activities and skills of a collection of enterprises, namely Belgian and French open source companies. In order to avoid manual annotation through visual analysis of the websites' content, a tool chain was developed to collect the content of websites and extract the important terms. Standard software libraries were identified, allowing to clean up HTML documents and to perform the part-of-speech tagging process used for extracting terminology. This procedure is supplemented by the extraction and the recognition of named entities. The terms extracted in the HTML pages of a company website were then merged and filtered and a circular tags cloud was generated. This presentation facilitates the identification of important terms, commonly referred to as activities and technologies supported by the company. Several changes are planned for this prototype, including, in particular, the extension to the texts in French, the association of extracted terms to the vocabulary of a classification scheme and the automatic generation of dashboards to facilitate the monitoring of the evolution of the industrial sector.**

Keywords—terminology extraction; named entities; NLP; tag cloud; market analysis

I. INTRODUCTION

As part of a market analysis process, a business directory needs to be maintained. Each entry is associated with a set of keywords to characterize the activities, technologies and software supported by the companies. These keywords are determined by the communication implemented by the company on its website. They are used to find specialized providers, to explore the market from a tag cloud and generate dashboards to compare the relative weight of technologies supported by the suppliers. This process to explore, annotate and update metadata is time-consuming. Research was therefore undertaken to accelerate and automate this task by establishing an information retrieval system, preferably based on standard tools.

This paper is organized in three parts. First, a state of the art on techniques and tools used to solve our problem was carried out. How to extract keywords from the content of a Web site will first be analyzed. This step will focus on two distinct problems; the conversion from HTML documents to raw text and the extraction of keywords to qualify business activity. A state of the art on techniques used to extract terminology and named entities will therefore be presented.

How to present the results of the extraction in order for the main topics of the company website to be understood will also be covered, followed by a presentation of results of a first implementation of a tool that extracts and formats keywords from a website. To conclude, possible improvements to this system will be discussed.

II. BACKGROUND

The metadata extraction process can be divided into three stages [7]:

1) *the conversion and the standardization of source files,*
2) *the part-of-speech tagging (POST),*
3) *the extraction of metadata.*

This process must be followed by the visualization of the extracted metadata.

B. Conversion of source files

The conversion of source files (in this case, HTML documents) is important because it determines the quality of the next step designed for part-of-speech tagging. Indeed the accuracy of the part-of-speech tagging influences the accuracy of the retrieval algorithms [22]. In practice, part-of-speech taggers often malfunction with data from the Web for various different reasons [2, 10]. Following the cleanup of HTML documents, the text entered into the tool may end up containing parasitic features, such as textual elements belonging to the menus, scripts, style sheets and footers. Additionally, the text may not meet the standards of written English (e.g. spelling errors, grammatical errors and specific writing styles) and may need to be standardized [9]. The taggers are also designed for a language (or set of languages) hence the need for prior configuration of the language of the document. Several approaches are possible for cleaning up documents.

A first approach is based on general tools for a crop (in French: "*détourage*") of the document. The term "crop" is proposed by Dutrey *et al.* by analogy with image processing. This means the separation of text and code in the context of digital structured or semi-structured documentation and/or separation within the textual content between relevant and irrelevant text [9]. Regarding HTML, Boilerpipe (refer to http://code.google.com/p/boilerpipe/) is a reference tool which enables an automatic cleanup of content pages and is distinguished by its good performance of low calculation time, ease of use and high accuracy [11].

A second approach relies on the special characteristics of the analyzed documents to extract content with better precision [17]. The use of reverse engineering tools for Web pages is possible when implementing this approach. These tools allow the content to be targeted more precisely and a first layer of semantics to be added. They may require writing extraction rules in the HTML document and may need to be distinguished by their ability to generate these rules semi-automatically [16]. However, they often use technologies such as XPath or XQuery, and can suffer from a lack of markup validity of HTML Web documents [7].

The conversion can be followed by normalizing a text, in other words, correcting spelling errors, deleting extra spaces, homogenizing punctuation, etc.

C. Part-of-speech Tagging

The part-of-speech tagging is a process of combining the words in a text and their grammatical function (e.g. noun, verb, etc.) based on lexical and contextual information [8]. Five criteria are used to select a part-of-speech software program, such as product support, license, available languages, accuracy and processing speed [1, 10, 20]. POST tools are widely available for English, but support for other languages is often poor. Various studies can assist in choosing a POST tool [9, 13, 20, 21].

The creation of a tagger for a given language requires detailed knowledge of the language and an important preliminary annotation. Unsupervised part-of-speech taggers could help overcome these constraints [3].

D. Extraction of Metadata

The extraction of metadata involves terminology extraction techniques which operate by extracting collocations, in other words, for example, word-pairs or word-triplets. This step is the result of the part-of-speech tagging. The extractor retains collocations such as Noun-Noun, Noun-Adjective, Noun-Preposition-Noun, etc. These collocations should then be filtered. According to Zhang *et al.*, the performance of filters depends on the type of document source (specialized or more general) [22]. The authors do not recommend the simple filtering of low-frequency words and encourage testing other related filters, for example, measures made on taken on sets of documents. Some systems only retain word-pairs and word-triplets. However, single terms should not be overlooked as they can be, in some areas, significant (e.g. gene names).

These terms may be extended by named entities via a specific extraction process. The notion of named entity refers to a unique and concrete entity, belonging to a specific domain [14]. In practice, it covers proper nouns, times and amounts. The Message Understanding Conferences (MUC) lecture series proposed a categorization with three categories and seven subcategories: Named Entities / ENAMEX (organization, location and person), Temporal Expressions / TIMEX (date and time) and Number Expressions / NUMEX (money and percent). Finer categorizations also exist. Sekine and Nobata offer 200 categories [18]. Systems exceeding 2000 categories also exist. They are used for the implementation of semantic labeling [6]. The problem of language support also arises with regard to the tools used for named entity recognition.

E. Visualization of Metadata

The principle of tag cloud can be used for the visualization of extracted metadata. The tag clouds are suitable for the exploration of content and the research of resources associated with wider research [5]. Lohmann, Ziegler and Tetzlaff studied the impact of the choice of format for tag clouds [12]. This study compares the presentation of tags in several forms: as a list arranged in alphabetical order (without changing the visual properties), as a list arranged in alphabetical order (with a change of visual properties), as a circular cloud (with the most popular tags in the center) and as a clustered cloud (tags belonging to the same theme put together). Searching for keywords in an alphabetical list is more efficient than searching for them in alphabetically arranged tags cloud. However, the latter is most suitable for identifying popular tags. Searching for popular tags in a circular cloud is easier as is searching within a clustered tag cloud for theme-specific tags.

III. Implementation

The tests were carried out on a set of websites from Belgian and French companies specializing in free and open source software. As regards the selection of tools, the use of free and open source software was the preferred choice. See figure 1 for the implementation scheme.

Wget, the free software, was used to recover the website content and store it locally (refer to http://www.gnu.org/software/wget/). By following the hyperlinks, this software allows recursive crawling. It also stores the website locally and retains its original structure. It is therefore easy to refine treatments to be carried out on the original content by using the local copy. The conversion of HTML documents into plain text is supported by Boilerpipe. This software also cleans up documents without requiring a specific configuration for each website, and retains only the useful content of the HTML document.

The tools for the part-of-speech tagging and the extraction of named entities are common in English but less common in French. Therefore, there was initially access to websites in English. However, it was noticed that websites in English could contain pages (or fragments of pages) written in other languages. For example, a site that was recovered stored different language versions in containers, in other words, <DIV> tags in HTML, of which the visibility was selected according to the desired language. A language test was introduced before the part-of-speech tagging process. The language is detected by the Java language-detection library (refer to http://code.google.com/p/language-detection/). The contents are filtered by threshold on the score of the first detected language. According to the selection criteria presented in the state of the art (support, license, available languages, accuracy and processing speed), the software OpenNLP was used (refer to http://opennlp.sourceforge.net). OpenNLP is a project supported and maintained by the Apache Software Foundation. It is published under the Apache License, a permissive free license, facilitating integration into development covered by different licenses (it is noted, however, that there is an inconsistency between the Apache v2 license and the GPL v2 license). Multiple languages are available, including English, Spanish and Dutch. OpenNLP

boasts a good reputation for accuracy and processing speed (e.g.: [4], [20] and [21]).

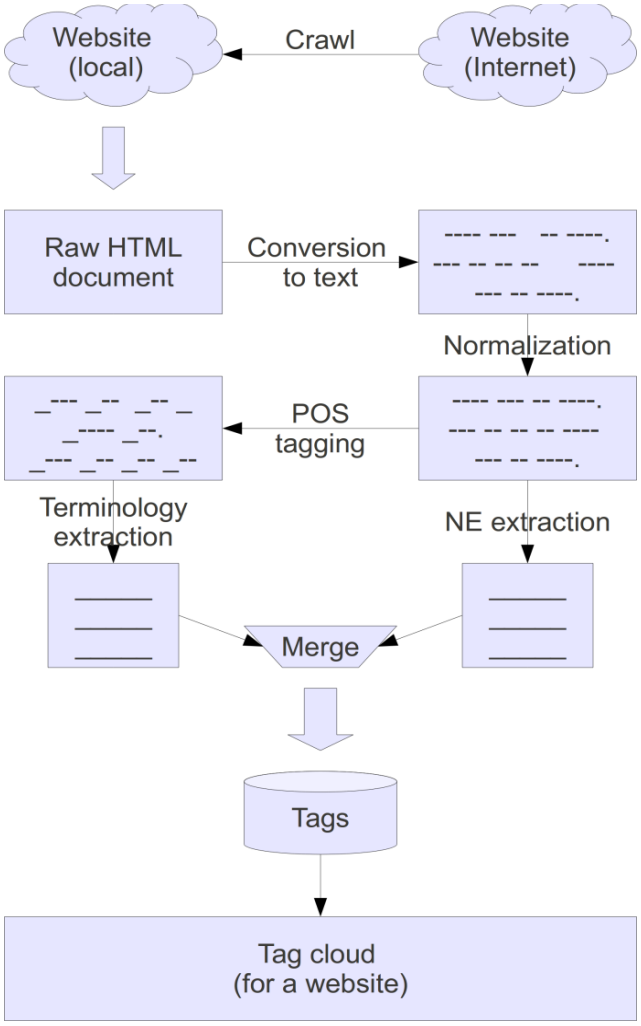

Fig. 1. Implementation scheme.

The extraction works by selecting and filtering collocations such as Noun-Noun, Noun-Adjective, Noun-Preposition-Noun, etc. Single terms were chosen if they corresponded to a proper noun, which include, typically, company names, software names, standards names, etc.

First, a minimum frequency was chosen for filtering the extracted terms. This approach remains common, yet it is criticized since it reduces the recall [22]. While the recall is not a criticism in itself and can be improved at a later stage, the principle objective is to quickly view the most common words. OpenNLP also provides the functions needed for the extraction of named entities in English, Spanish and Dutch. It was used to extract the person named entities.

Two types of formatting, for the retained terms were implemented after the filtering stage. This demonstrated the principle of the tag cloud. The first presentation involves putting the tags in a tag cloud, with an alphabetical list and highlighting the most important terms (see Figure 2).

The second presentation classifies the terms from most the important to the least important, then formats them in a circular tags cloud (see Figure 3). The circular tags cloud seems well suited to quickly view the people, themes and activities that are important for a company.

IV. Conclusion

This first prototype has validated the principle of the automated annotation of a company directory based on the content of Web sites. It also instigates interesting perspectives regarding the development of sectorial market analysis activities, the basis of this project.

This project has, however, highlighted the difficulty of having sustainable and freely available tools for part-of-speech tagging, but particularly for the extraction of named entities, as in the case of texts in French.

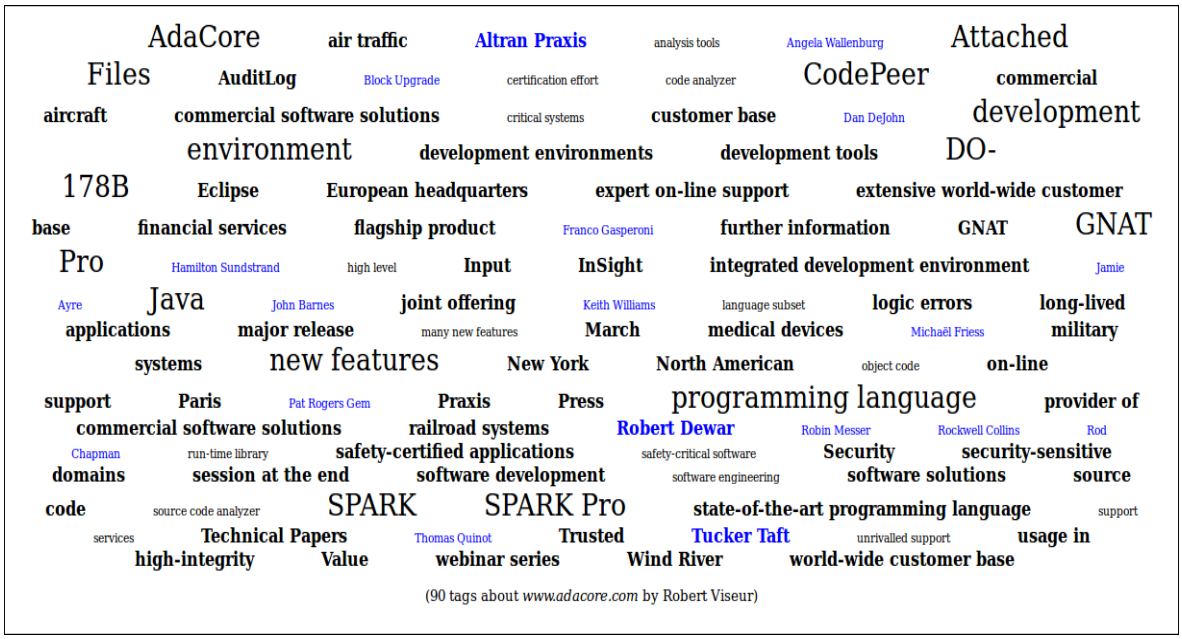

Fig. 2. Classic tag cloud.

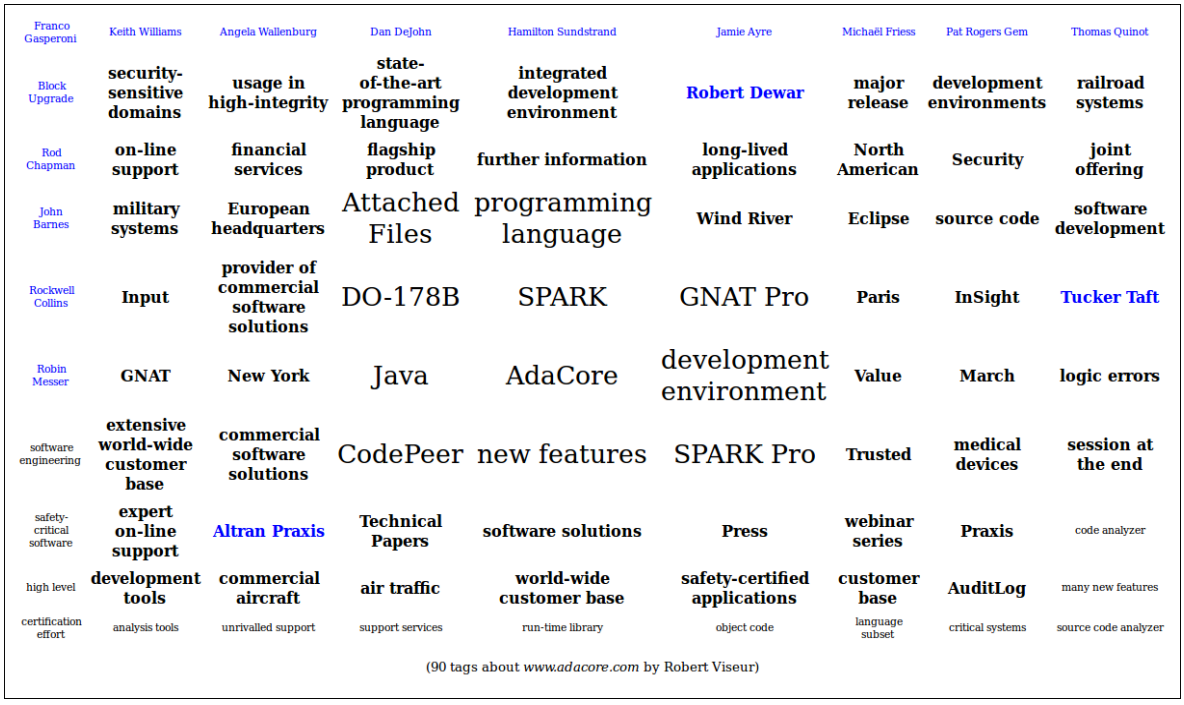

Fig. 3. Circular tag cloud.

V. FUTURE WORKS

Only the English language is currently taken into account. Considering the French language is a must due to the high number of French companies in the industrial sector that shall be studied. Adding a French model to OpenNLP is possible but requires specific skills. Using other part-of-speech software is also possible. Unsupos (refer to http://wortschatz.uni-leipzig.de/~cbiemann/software/unsupos.html) is a possibility in this case since it supports English, French, Dutch, German and Italian, as is Stanford Log-linear Part-of-Speech Tagger (refer to http://nlp.stanford.edu/software/tagger.shtml) which supports the French language in particular (since January 2012) as well as English and German [3, 19]. For the named entity extraction and recognition, a first evaluation of TagEN software was performed on a real corpus (French web content that was previously cleaned), with encouraging results for dates, locations and people, but unsatisfactory for organizations names (low recall and partial extraction for multi-word expressions).

As with folksonomies, terms extracted from a set of Web sites are subject to change depending on the communication policy of a company and the person in charge. This approach lacks controlled vocabulary. The system could therefore benefit from the establishment of a link between the extracted terms and vocabulary from a taxonomy (see [6] and [15]). Using a thesaurus (a list of controlled terms enriched by pre-defined associative relationships) or an ontology (a descriptive knowledge model based on concepts with types, properties and relations) instead of a taxonomy (a list of controlled terms organized hierarchically) would potentially lead to a more in-depth analysis of the possibilities of automatic generation of dashboards (e.g. dashboards by type of software). This may require the creation of a classification scheme, possibly using already existing tools (e.g. dictionaries, Wikipedia/DBpedia, etc.).

The ability to automatically extract common terms may allow evolutions in technology adoption to be investigated and developed. Pirolli offers this type of development for tag clouds [14]. The emergence of new tags in folksonomy can in fact show a phenomenon of craze or disinterest.

We do not use tags from folksonomy but terms extracted from websites' content (i.e. outcome of corporate communication), which can lead to the same kind of phenomenon. The terms monitored over time should allow the commercial life cycle of a technology in an industrial sector to be visualized.

REFERENCES

[1] J. Asmussen, "Survey of POS taggers", DK-CLARIN WP 2.1 Technical Report, Final version of August 19, 2011.

[2] B. Baroni, F. Chantree, A. Kilgarriff, and S. Sharoff, "CleanEval: a competition for cleaning webpages", Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008.

[3] C. Biemann, "Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering", Proceedings of the COLING/ACL-06 Student Research Workshop 2006, Sydney, Australia.

[4] F. Boudin and N. Hernandez, "Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank", Actes de la conférence conjointe JEP-TALN-RECITAL 2012, Vol. 2, pp. 281–291.

[5] M. Cardew-Hall and J. Sinclair, "The folksonomy tag cloud: Is it useful", Journal of Information Science, vol. 34, no. 1, 2008, pp. 14-29.

[6] E. Charton, M. Gagnon, and B. Ozell, "Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique", TALN 2010, Montréal, 19–23 juillet 2010.

[7] S. Chen, D. Hong, and V. Y. Shen, "An experimental study on validation problems with existing html webpages". In International Conference on Internet Computing ICOMP 2005, pp. 373-379.

[8] M. Dieye, M.R. Doulache, M. Floussi, J. Chabalier, I. Mougenot, and M. Roche, "Construction d'un dictionnaire multilingue de biodiversité à partir de dires d'experts". In Proceedings of InforSID 2012, May 29-31, Montpellier, France.

[9] C. Dutrey, A. Peradotto, and C. Clavel, "Analyse de forums de discussion pour la relation clients : du Text Mining au Web Content Mining", Actes JADT'2012, Liège (Belgique), 13-15 juin 2012.

[10] S. Evert, and E. Giesbrecht, "Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus.", Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, Spain, 2009.

[11] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate Detection using Shallow Text Features", Third ACM International Conference on Web Search and Data Mining (WSDM 2010), New York City, USA.

[12] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration", Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT '09), pp. 392-404.

[13] M. Marrero, S. Sánchez-Cuadrado, J. Morato, G. Andreadakis, "Evaluation of Named Entity Extraction Systems, Research", Computing Science, Vol. 41, 2009, pp. 47-58.

[14] C. Martineau, E. Tolone, and S. Voyatzi, "Les Entités Nommées : usage et degrés de précision et de désambiguïsation". In Catherine Camugli Gallardo, Matthieu Constant, and Anne Dister, editors, Actes du 26ème Colloque international sur le Lexique et la Grammaire (LGC'07), Bonifacio, France, Octobre 2007, pp. 105-112.

[15] F. Pirolli, "Apports des folksonomies dans le cadre d'un processus de veille : vers la prise en compte des spécificités informationnelles", ISKO-France, Lyon, France, 2009.

[16] B.A. Ribeiro-Neto, A.S. da Silva, J.S. Teixeira, "A brief survey of web data extraction tools", ACM SIGMOD Record, Volume 31 Issue 2, June 2002, pp.84-93.

[17] M. Roche, T. Heitz, O. Matte-Tailliez, and Y. Kodratoff, "EXIT: un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés", JADT'04, Mars 2004, Belgique.

[18] S. Sekine, and C. Nobata, "Definition, dictionaries and tagger of Extended Named Entity hierarchy", Proceedings of LREC, 2004.

[19] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", Proceedings of HLT-NAACL 2003, pp. 252-259.

[20] M. Wilkens, "Evaluating POS Taggers: Speed", November 8, 2008, http://mattwilkens.com (read: January 31, 2014).

[21] M. Wilkens, "Evaluating POS Taggers: Coda", February 9, 2009, http://mattwilkens.com (read: January 31, 2014).

[22] Z. Zhang, J. Iria, C. Brewster, and F. Ciravegna, "A Comparative Evaluation of Term Recognition Algorithms". In Proceedings of The sixth international conference on Language Resources and Evaluation (LREC 2008), May 28-31, 2008, Marrakech, Morocco.