CONCEPTS EXTRACTION BASED ON HTML DOCUMENTS STRUCTURE

Rim Zarrad[1], Narjes Doggaz[2] and Ezzeddine Zagrouba[1]
[1] *RIADI Laboratory, Team of research SIIVA, High Institute of Computer Science , University El Manar, Tunis, Tunisia*
[2] *URPAH laboratory, Faculty of Sciences of Tunis, University of Tunis El Manar, Tunis, Tunisia*

Keywords: Document Structure, Keyword Extraction, Title, Web.

Abstract: The traditional methods to acquire automatically the ontology concepts from a textual corpus often privilege the analysis of the text itself, whether they are based on a statistical or linguistic approach. In this paper, we extend these methods by considering the document structure which provides interesting information on the significances contained in the texts. Our approach focuses on the structure of the HTML documents in order to extract the most relevant concepts of a given field. The candidate terms are extracted and filtered by analyzing their occurrences in the titles and in the links belonging to the documents and by considering the used styles.

1 INTRODUCTION

Due to the rise of the Web and the need to have structured knowledge, a big part of research concentrates on the formalization of ontologies (Berners-Lee et *al*., 2001). The methods dealing with ontologies building are mainly based on the techniques of natural language processing. These methods can be clustered in three groups according to the chosen technique: the methods mainly based on the distribution of the linguistic units (Bourrigault et *al*., 2005), the ones based on the statistical approaches (Velardi, 2002) and those based on the semantic and/or the linguistic approaches (Séguéla, 1999, Morin, 1999). The methods which are based on the distributional processing distinguish the concepts of the ontology and organize them in structured systems reflecting a conceptual hierarchy. The statistical methods tend to extract conceptual links which are difficult to dissociate without recourse to an expert of the field. The linguistic approach does not consider the corpus in a comprehensive manner but locally. It is based on the properties of the language which define those of the objects of the field.

In this paper, we propose an approach which differs from the traditional approaches since it uses information on the document structure to extract relevant information. Our approach studies each material form in the text (title, style, and hyperlink) in order to extract the concepts of the ontology. This paper is organized as follows. In section 2, we present the related works which study the document structure in order to build ontologies. In section 3, we present our approach by giving the different steps to extract the concepts of the ontology. The concept refiner step is described in details in section 4. We give in section 5 some experimental results to evaluate the proposed approach. Finally, we conclude our paper by a summary and some perspectives.

2 RELATED WORKS

The traditional methods of building ontology and concept extraction from texts often privilege the analysis of the text itself, whether they are based on a statistical or linguistic approach. New approaches appeared in the last decade, they exploit the material formatting of the texts in order to build conceptual models or ontologies. In fact, the visual properties of texts are not just an ornamentation of the text but constitute an important component implied in the significance. The approach described in (Karoui et *al.*, 2004) studies the structure of Web pages to build a database table. It uses a clustering procedure to extract the concepts of the ontology. It is based on the TF*IDF measure (Joachims, 1997) which is a weighting method often used in the field of

Zarrad R., Doggaz N. and Zagrouba E..
CONCEPTS EXTRACTION BASED ON HTML DOCUMENTS STRUCTURE.
DOI: 10.5220/0003748305030506
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 503-506
ISBN: 978-989-8425-95-9

knowledge extraction.

$$TFi = ni \ / \Sigma j \ n \ j \quad (1)$$

where n_i is the number of occurrences of the i^{th} candidate term (CT) in the document.

$$IDF = \log (|D| / |\{dj / ti \in dj \}|) \quad (2)$$

Where d_j is the j^{th} document of the corpus and |D| is the total number of documents of the corpus.

Among the methods based on the document structure, several ones focuse on the titles belonging to the document. The method presented by (Hazman et *al*., 2009) uses the hierarchical structure of the HTML headings for identifying new concepts and their taxonomical relationship between seed concepts and between each other. An analysis (Lopez et *al.*, 2010) was made on a corpus of web documents. The results are interesting when analyzing terms which appear in the titles. The titles play an important role in the level of the material organization of the text: they segment, make taxonomies, and provide a denomination for the obtained segments. Moreover, it is obvious that the text snippets which are strongly marked from a typographical point of view are more important than those which have not such marking. The cognitive psychologists proved by experiments that the styles used in the text have a big impact on its comprehension.

3 PROPOSED APPROACH

Our objective is to extend the traditional approaches which aim to extract the concepts and the links between them by considering the text structure in the corpus. The proposed approach described below is based on the study of several web sites in order to find the concepts of the field.

3.1 The Corpus Constitution and Pre-processor

The corpus must be representative of the field for which we try to build ontology. Our approach is based on a corpus of HTML documents collected using Google Web Search API. It is a library offered by the Google programmers in order to extract the results of a request. The corpus pre-processor is performed in three steps:

Tagger. Before being analyzed, the corpus is treated by the Tree tagger (Schmid, 1994). This tagger associates each instance of a word with its grammatical category and with its canonical form.

Parser. We use HtmlParser 1.6 which is a free library for the extraction and the text processing of the corpus generated from the web. It extracts data from the tags of the HTML documents.

Stop-Lists. A general stop-list is used to locate the blank words which occur in the corpus (articles, prepositions…). Another stop-list is created in order to check if a title has an empty informational content (introduction, conclusion…).

3.2 Extraction of Candidate Terms

Our objective is to detect within the corpus the interesting terms which are candidates to represent the concepts of ontology. They are linguistic units which qualify an object of the real-world. We extract two types of syntagms according to their structures. Firstly, we extract syntagms composed by only one word, they can be either a "Name" or a "Named Entity". Secondly, we consider syntagms containing two words: they have as syntactic structure the sequence "Adjective Name". The CT which are extracted will be then filtered according to their appearance in the titles, styles and hyperlinks used in the corpus.

4 THE CONCEPTS REFINER

To consider the relevance of a syntagms S in the corpus, we was inspired by the measure TF*IDF (Joachims, 1997). Indeed, we propose a new measure CR*IDF (CR_IDF) where CR, the Corpus Relevance of a candidate term S, corresponds to its relevance in all the documents of the corpus. The candidate terms will be then filtered chosing those having a value CR_IDF higher than a given threshold. The selected candidate terms represent the relevant terms or concepts of the field.

Definition 1: Given a corpus C = {d_1, d_2,…, d_m} of m HTML documents (d_j), we define the Corpus Relevance CR of each syntagm S as the sum of its Normalized Relevance NR in all the documents of the corpus.

$$CR(S) = \sum_{j=1}^{m} NR(S, d_j) = \sum_{j=1}^{m} \left(\frac{R(S, d_j)}{\sum_{k=1}^{n_s(j)} R(S_k, d_j)} \right) \quad (3)$$

where $n_s(j)$ is the number of syntagms in the document d_j and R(S,dj)is the relevance of the syntagm S in the document dj. The denominator is the sum of the relevance of all the syntagms in this document. The normalized relevance of a syntagm S in the document d_j is used to avoid the problems

related to the length of the document.

Definition 2: The relevance R of a syntagm S in a document d_j is computed by:

$$R(S, d_j) = R_{Title}(S) + R_{Style}(S) \quad (4)$$

where $R_{Title}(S)$ and $R_{Style}(S)$ are the Title and Style relevances of the syntagm S in the document.

In the rest of this section, we consider the document d_j of the corpus and we define the relevance of each syntagm in the titles and the styles of the document.

4.1 Title Relevance

The titles and the subtitles generally contain the relevant terms of the field. It would be judicious, then, to consider them for the selection of concepts. Moreover, we note that the syntagm relevance in a given title strongly depends on the length of this title. Let title length be the number of syntagms which it contains. Thus more the number of syntagms in a title is low, more the syntagm relevance in this title is high. To illustrate this idea, let consider the two following titles "La spectroscopie" (spectroscopy) and "Calendriers et instruments de mesure du temps" (schedule and tools of the time measurement) extracted from a French astronomy Wikipedia document. In the first title, the obtained syntagm (spectroscopy) is more representative than the four syntagms composing the second title which are less significant (schedule, tool, time and measurement).

Definition 3: We note $T_i = \left(t_1^i, ..., t_{n_i}^i\right)$ the set of the titles appearing in the level i of a HTML document where n_i is the number of titles of the level i. The relevance r_s of a syntagm S which appears in the title t_k^i is computed as follows:

$$r_s(t_k^i) = \frac{1}{l(t_k^i)} \quad (5)$$

where $l(t_k^i)$ corresponds to the length of the title t_k^i (i.e. the number of syntagms in this title).

Applying this formula to the two titles given in the example above, we have $r_{spectroscopy} = 1$, $r_{schedule} = 0.25$, $r_{tool} = 0.25$, $r_{time} = 0.25$, $r_{measurement} = 0.25$. We can note that these results reinforce our intuition.

Definition 4: The relevance $R_s(i)$ of a syntagm S in all the titles appearing in the level i of a HTML document is computed as follows:

$$R_s(i) = \sum_{k=1}^{n_i} r_s(t_k^i) \quad (6)$$

where n_i is the number of titles of the level i.
Now, we can calculate the title relevance of a syntagm in a document.

Definition 5: The Title relevance $R_{Title}(S)$ of a syntagm S corresponds to its relevance in all the titles of the document. It is computed as follows:

$$R_{Title}(S) = \sum_{i=1}^{6} \frac{1}{i} * R_s(i) = \sum_{i=1}^{6} \sum_{k=1}^{n_i} (i * l(t_k^i))^{-1} \quad (7)$$

We consider that more the depth level of a title is low, more the syntagm relevance in this title is high. Indeed, we give the weight 1/i to each level i.

4.2 Style and Hyperlink Frequency

The words having a particular style (bold, italic) or belonging to the hyperlinks of the documents are easily detectable in the corpus by analyzing the tags of HTML documents. For each CT extracted, we define four coefficients: f_{Bold} (S) which corresponds to the Bold style frequency of a syntagm S in a HTML document, f_{Italic} (S) which consists in its Italic style frequency, f_{url} (S) which corresponds to its hyperlink frequency and f_{Rest} (S) corresponding to the appearance frequency in the rest of the document.

Definition 6: Let be Format = {Bold, Italic, URL, Rest}. We define the style relevance $R_{Style}(S)$ of a syntagm S as follows:

$$R_{Style}(S) = \sum_{Style \in Format} w_{Style} * f_{Style}(S) \quad (8)$$

where w_{Style} is a weight given to each style and it was fixed experimentally.

5 EXPERIMENTATION

We have evaluated the contribution of our method on a corpus of 23 documents of Wikipedia which are related to the astronomy field. The total number of candidate terms extracted from these documents is 2801. After the refinement step, only 134 concepts (105 one-word concepts and 29 two-word concepts) remained. The refinement of these CT is realized by removing the syntagms belonging to a limited number of documents and those having a CR*IDF value less than a constant threshold. To validate the extracted concepts, we verify their belonging to a

Table 1: Evaluation of the obtained concepts using CR, CR_IDF and TF_IDF mesures.

Measure / Extracted concepts		CR		CR_IDF		TF_IDF	
Type	**Number**	**Affirmed Concepts**	**Precision (%)**	**Affirmed Concepts**	**Precision (%)**	**Affirmed Concepts**	**Precision (%)**
One-word Concept	105	44	41.90%	71	67.62%	61	58.09%
Two-word Concept	29	16	55.17%	18	62.07%	9	31.03%
Total	134	60	44.78%	89	66.42%	70	52.24 %

specialized dictionary in astronomy and space (Cotardière and Penot, 1999). We have also fixed $w_{Bold}=3$, $w_{Italic}=2$, $w_{url}=2$ and $w_{rest}=1$.

In table 1, we present the results obtained using CR (Corpus Relevance) and CR_IDF measures. The CR_IDF measure gives more interesting results than CR ones. This can be explained by the fact that the use of IDF factor removes CT which are the less specific to the considered field. To evaluate our approach, we compare our results (CR_IDF) with the most used measure in the literature TF_IDF. We note that our CR_IDF measure which considers the documents structure gives better results than TF_IDF. These results show the impact of considering the documents structure in the extraction of concepts by giving the most relevant ones.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we have defined a strategy which extracts automatically the concepts of the ontology from a corpus of Web documents. This strategy is based on the study of the document structure by extracting the typographical titles, links and markings. Indeed, the structure of the documents provides interesting information on the significance contained in the texts. We can extend the work performed by analyzing the hierarchy of the titles in each document in order to extract the hierarchical links to lead to ontology. The text can be then apprehended not like a linear succession of blocks of various natures, but like a structure of elements of high level which include other elements.

REFERENCES

Berners-Lee, T. Hendler, J., Lassila, O, 2001. The Semantic Web. Scientific American, pp. 28–37

Bourrigault, D., Fabre, C., Frérot, C., Jacques, M. P., Ozdowska, S., 2005. Syntex, analyseur syntaxique de corpus. *In: Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, pp. 17-20, Dourdan, France

Cotardière, P., Penot, J. P., 1999. Dictionnaire de l'Astronomie et de l'Espace. eds. Larousse-Bordas

Hazman, M., El-Beltagy, S.R., Rafaa, A., 2009. Ontology Learning from Domain Specific Web Documents. *International Journal of Metadata, Semantics and Ontologies*, Vol 4, Number 1-2, pp. 24-33

Joachims, T., 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *In: Proc. 14th International Conference on Machine Learning*, pp. 143- 151, Morgan Kaufmann

Karoui, L., Aufaure, M. and Bennacer, N., 2004. Ontology Discovery from Web Pages: Application to Tourism. *In: Knowledge Discovery and Ontologies Workshop at ECML/PKDD*

Lopez, C., Prince, V., Roche, M., 2010. Automatic Titling of Electronic Documents with Noun Phrase Extraction. *In: Proceedings of Soft Computing and Pattern Recognition, SOCPAR'10, pp. 168-171, Paris, France*

Morin, E., 1999. Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus. *In: Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, pp 268-278

Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *In: Proceedings of the International Conference on New Methods in Language Processing*

Séguéla, P., 1999. Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Revue Terminologies*, Number 19, pp. 52-61

Velardi, P., Fabriani, P. and Missikoff, M. , 2002. Using text processing techniques to automatically enrich a domain ontology, *In Proceedings of the ACM Conference on Formal Ontologies and Information Systems*, pp 270-284