A framework for retrieving conceptual knowledge from Web pages

Nacéra Bennacer, Lobna Karoui

Ecole Supérieure d'Electricité (Supélec),
Plateau de Moulon 3 rue Joliot Curie, 91192 Gif-sur-Yvette cedex, France
{Nacera.Bennacer,Lobna.Karoui}@Supelec.fr

Abstract. Ontologies provide a common layer which plays a major role in supporting information exchange and sharing. Their proliferation relies strongly on the automation of ontology building, integration and deployment processes. In this paper we introduce an integrated framework involving different and complementary dimensions to drive the (semi) automatic acquisition conceptual knowledge process from HTML Web pages. Our approach takes advantage from both structural and linguistic HTML document characteristics and is based on an incremental evaluation by the user of the conceptual quality.

1 Introduction

Ontologies provide a common layer which plays a major role in supporting information exchange and sharing by extending syntactic interoperability to semantic interoperability in the semantic Web. The proliferation of ontologies determines the success of the semantic Web and it relies strongly on the ontology building, integration and deployment processes. The major problem to be faced is time-consuming construction of various ontology for various domains and applications thus moving towards automation of ontology building seems to be the key of this problem. A variety of works found in the literature originating from diverse communities and entailing complementary fields ranging from data mining, databases, software engineering, and linguistics has actually researched and practiced techniques for solving parts of the overall problem. Defining the process that retrieves a set of concepts of a domain and their taxonomic relations is a fragment and a beginning of a more ambitious and complex task which is the building of ontologies.

Most works have investigated various issues of ontology building such as methodology frameworks [1] et [2] and automation aspects. They are distinguished essentially by:

- their input types of unstructured (free text for example), semi-structured (HTML and XML documents for example) and fully structured such as database schemas.
- the use of restricted vocabulary such as technical documents,

– the use of a priori knowledge like a taxonomy (wordnet for example) or an existent ontology

Automatic processes that derive knowledge from texts belong to two kind of approaches, radically different, data mining and linguistic approach with a lot of variations between two extremes. Central to these approaches is the surface analysis of texts based on the distributional hypothesis which assumes that terms are similar because they share similar contexts. Purely data mining approaches can be applied to large corpus whatever their domain and obtain clusters of words by exploiting lexical knowledge and by analyzing importance of words in a corpus, their frequency and their co-occurrences. On other hand, linguistic approaches are more syntactical and characterize how words are used. They are designed to work on a specialized language with a limited and well defined vocabulary.

In this paper we introduce a unified framework to the automatic acquisition of conceptual knowledge from HTML pages. Indeed, the abundance and the importance of HTML pages as a principal source of rich information available on the Web is an undeniable fact particularly with the continuous and rapid expansion of this one. Moreover, Web pages were originally designed to be human-readable thus they contain semantic information hidden in the structure and the presentation of a HTML document. Mining this resource to provide mechanized ways to derive concepts and their relations seems to be both a promising and a challenging issue. Our approach encompasses several dimensions simultaneously and is based on an incremental evaluation of the conceptual quality. It takes advantage from both structural HTML document and linguistic characteristics.

The remainder of the paper is organized as follows: section 2 describes the different modules of our framework, section 3 presents the related work and section 4 concludes on our work and its continuity.

2 Our Framework

Our framework is composed principally of processing and structuring module, analysis and characterization module and extraction conceptual knowledge module. The purpose in the first module is to transform HTML web pages into structured data represented by a relational table. The second module enriches this relational representation by characterizing its structural and linguistic features in order to determine precisely the context of a term and its vicinity (see Figure 1).

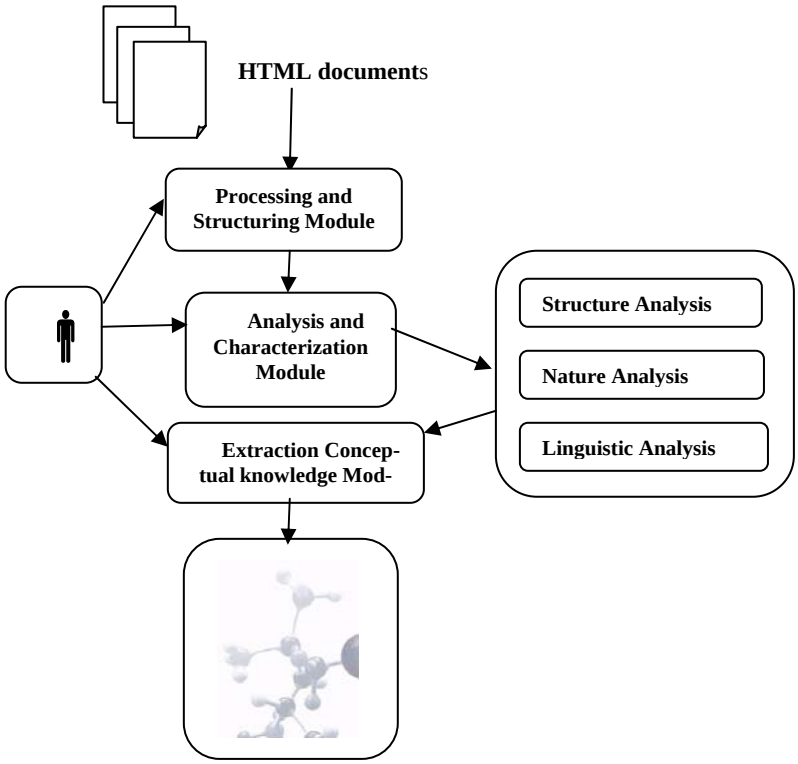

Figure 1 : Our Framework Architecture

2.1 Processing and Structuring Module

As mentioned above, our proposal is intended for HTML web pages which are rich by their textual content and its associated structure and presentation using block-level and inline markups. Block-level markups such as <h1> markup to indicate an important heading, <h2> markup to indicate a slightly less important heading, <p> markup to indicate a paragraph, <ul>, <li> and <ol> markups to define an unordered and nested lists, <dl>markup to introduce a list of definitions, and markup for link with other pages, typically contain inline or other block-level markups. While inline markups such as <b>, <i>, <big> may only contain text and inline elements to emphasize important words. The purpose in this module is to constitute a corpus, to process it by keeping important markups related to text and defining new ones.

Corpus constitution and processing

Our corpus is constituted of a set of HTML pages selected from a set of sites concerning a specific domain. For our experiences we use a corpus constituted from a set of

Figure 2 : An Example of processed HTML document

HTML pages of French official tourism sites [3]. The corpus is firstly processed to keep only the text associated to a set of markups considered to be important to retrieve the most important terms. We define other tags to emphasize important terms like those titling an hyperlink by the key tag <TITLE_URL> or a check box by the key tag <CHOICE>. We use the key tag <KEYWORDS> to all elements of meta data associated to a document. We then proceed to correct some character coding in particular some words which are not correctly accentuated using Réacc tool for French text [4] (see Figure 2).

Corpus relational representation

Once the corpus processed, it is represented in database table which is defined by a set of attributes. The subset of attributes that are filled in this stage of study, allow to relate each term to its markup , its previous one and its ranking in its source document. This transformation keeps the whole corpus and the information related to the link between markups and the link between documents. Moreover this relational representation enables a backtracking to source document and performing sql requests suitable to have synthetic and targeted information about the corpus each time the user needs (see table below).

	term	grammatical_type	lemma	associated_tag	N°_associated_tag	previous_tag	N°_previous_tag
	héberg	Nom	héberger	<title>	1	<URL_page>	1
	agence	Nom	agence	<title>	1	<URL_page>	1
	héberg	Nom	héberger	<keyword>	1	<title>	1
	<a href	Nom	<unknow	<Lien_hypertexte>	1	<keyword>	1
	Votre	pronom	votre	<Titre_URL>	1	<Lien_hypertexte>	1
	héberg	Nom	héberger	<Titre_URL>	1	<Lien_hypertexte>	1
	Agence	Nom	agance		2		2
	immobi	adjectif	immobili		2		2
	grand	adjectif	grand	<b>	1	<titre_URL>	3
	domain	Nom	domaine	<b>	1	<titre_URL>	3

Table 1 : An Extract of relational representation of corpus

2.2 Analysis and Characterization Module

This module is composed of three complementary kinds of analysis in order to evaluate and to characterize structural, nature and linguistic corpus features.

Structure Analysis
This study allow to evaluate the structural features of the considered corpus such as markup diversity by computing markup frequency for each category, and associated term percentage. Structural patterns can also be discovered to determine markups that appear together like heading followed by paragraph or a list of sentences ended by a special character. This analysis propose to the user a set of structural patterns that allow to refine the term context definition by delimiting its vicinity and to choose the appropriate weighting measure for clustering procedure.

Nature Analysis
Our HTML pages are selected from sites concerning the same domain so this is not sufficient to say if the obtained corpus is homogeneous or if it covers all the domain. We must analysis the lexical information of our source and especially its distribution in the same space or between documents. To achieve this purpose, we use two kinds of methods. The first method is a factorial correspondence analysis [5] which provides a geometrical representation where the proximities between line point and column points represent statistical associations between profiles of the original table. It presents indicators on the quality representation and on the respective words and documents contributions to the construction of space axes. In our case, it permits to determine the diversity and homogeneity of our corpus The purpose here is to evaluate the term distribution which could be centered around one point or around a multitude of points. The second method is the TFIDF measure [6] which assigns a greater significance to terms that are good discriminators between documents. It compares how frequently a term appears in a document against the number of other documents which contain that term. This analysis can lead to change the corpus content by suppressing or adding HTML documents until we obtain homogeneous covering the considered domain.

Linguistic Analysis and Characterization
Morphological analysis is the identification of a word-stem from a full word-form and also the identification of the syntactic category of the stem. Inflectional morphology covers the variant forms of nouns, adjectives and verbs owing to changes in (Person, Number, Tense, gender) and derivational morphology is the formation of a new word of a different syntactic category. In our case, we use the TreeTagger tool [7] in order to assign a syntactic category and a stem to each term of our corpus. This information enriches the relational table by filling attributes related to linguistic characteristics. The purpose of syntactic analysis is to derive patterns based on different nominal group, verbal group, syntactic category (verb, noun, adjective, adverb, etc.). We make use of syntactic dependencies to refine the definition of term context (its

vicinity) and its semantic relation with other terms such as nouns appearing with the same verb. For making this analysis, we use Syntex tool [8].

2.3 Extraction Conceptual knowledge Module

Clustering methods is characterized by the use of a similarity or distance measure in order to compute the pairwise similarity or distance between vectors corresponding to two terms in order to decide if they can be clustered or not. Some examples of similarity measure are : cosine, Euclidian distance, jaccard, etc. The user can compare the results obtained by applying different similarity measures.

To weight the significance of a given term pair we combine two types of measures: co-occurrence in a structural context and co-occurrence in a syntactic context. The first one is defined by the existing links between two HTML markups like <h1> → <p>, <caption>→ <td>, <dt> → <dd>, <TITLE_URL> → headings of a part of document, <TITLE_URL> → headings of the referenced document, <TITLE> → headings of the document. If two terms are emphasized by the same block level tag (example 1) the context is delimited by the tag and their co-occurrence is computed in this context. If two terms are emphasized by different tags that are related structurally their co-occurrence is computed regarding this link in this context. For example, when we find two terms in the same phrase (or syntactic pattern)), we can speak about co-occurrence between these two words in a context phrase. In our study, the notion of context depend on analysis results.

Example 1:
<H1> museum visits </H1>
Example 2:
<TITLE> Visit florida | experience florida **attractions**, florida **entertainment** and florida **activities** </TITLE>
<KEYWORDS> *** </KEYWORDS> <HYPERLINK> *** <TITLE_URL> ***
<H1> classic florida **attractions** </H1>
<P> Florida's early roadside attractions strove to satiate America's appetite for the eccentric with rare animals, flamboyant gardens, crowd-thrilling acts and human oddi-ties </P>
In the example 2, if we consider term "attractions" we can found terms occurring in the same tag <TITLE> and in the tag <H1>. For the first case, the context is the tag himself <TITLE> but in the second case the context is two tags together (<TITLE> + <H1>).

The principles of a generic clustering procedure is defined by an initial hierarchy cluster obtained from key words tags corresponding to the most important terms. Leaf clusters are then refined by considering each co-occurrence terms in both structural and syntactic contexts. A tree is defined to represent markup hierarchy which is used to guide clustering procedure to iteratively consider two terms belonging to the considered hierarchy level. This iterative clustering allow the user to evaluate cluster at

each step. We experiment these heuristic refinement on a French corpus concerning tourism domain the first results seem to have promising issue.

3 Related Work

In this section, we discuss some work related to the automatic acquisition of taxonomies. Faure and Nédellec describe [9] a system called ASIUM where a cooperative conceptual clustering is applied to technical texts using syntactic parser to produce an acyclic conceptual graph of clusters. Basic clusters are formed by words that occur with the same verb after the same preposition.
Chalendar and Grau in [10] design a system called SVETLAN able to learn categories of nouns from texts taking into account the contextual use of words whatever their domain. Ciamiano and al [11] examine different clustering methods and provide a conceptual clustering method based on Formal Concept Analysis where the linguistic context of a term is defined using syntactic dependencies that it establishes as the head of a subject, of an object or of a complement with a verb.
In [12] the authors introduce a methodology for the maintenance of domain-specific taxonomies. It is embedded in a framework named Syndicate [13] which relies on two major kinds of knowledge grammatical knowledge for syntactic analysis and conceptual knowledge expressed in KL-One representation language.
Other approaches are developed for extracting relationships other than taxonomy relationship. In [14] Meadche and Staab propose a new approach to extend current one by focusing on discovery of non-taxonomic conceptual relations using the association rules algorithm. This approach doesn't identify the types of semantic relations that are discovered by the algorithm.
Bisson and al [15] present an interesting framework and a corresponding workbench Mo'K allowing users to design, compare and evaluate conceptual clustering methods to assist them in an ontology building task. Maedche and al [2] propose an ontology learning framework including ontology import, extraction, pruning, refinement and evaluation techniques implemented inText To Onto environment.

4 Conclusion

Building conceptual knowledge should encompass several dimensions simultaneously and should avoid any explicit specialized heuristics to drive the acquisition process. Our work aimed at providing an integrated framework including different and complementary analysis to enhance and refine the acquisition process. Structural and linguistic analysis allow to emerge information about concept characteristics such as structural patterns and syntactic dependencies. It is important to outline that the context definition varies according to the studied corpus. The relational representation of the whole corpus with the different analysis results facilitates the acquisition process. This one is carried out by clustering procedure which is applied iteratively on clusters obtained by weighting the co-occurrence term pairs according to structural and syn-

tactic context. The successively formed clusters are examined and validated by the user. We are implementing and experimenting the different analysis dimensions on a French corpus concerning tourism domain. The first results are interesting and allow to pursue this work to improve the acquisition process.

References

1. Gomez-Pérez, A., Fernandez-Lopez, M., Corcho, O. : Ontological Engineering, Springer. 2004.
2. A. Meadche and S. Staab : "Ontology learning for the semantic Web", IEEE journal on Intelligent Systems, Vol. 16, No. 2, 72-79, 2001.
3. L. Karoui, M-A. Aufaure, and N. Bennacer: « Ontology Discovery from Web Pages: Application to Tourism », ECML/PKDD 2004: Knowledge Discovery and Ontologies KDO-2004.
4. Simard, M., Deslauriers, A.: Real-time automatic insertion of accents in French text, Natural Language Engineering, Volume 7, no 2. Juin 2001. pp 143-165
5. Benzecri, J.-P. : L'analyse des correspondances. Dunod, 1973.
6. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proc. 14th International Conference on Machine Learning, Morgan Kaufmann, pages 143-151, 1997.
7. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. IMS-CL, Institut Für maschinelle Sprachverarbeitung, Universität Stuttgart, Germany, 1994.
8. Frérot, C., Bourigault, D. and Fabre, C. : Marier apprentissage endogène et ressour-ces exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in Revue t.a.l., 44-3, 2003
9. Faure, D., Nédellec, C., : A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In Paola Velardi, editor, LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications, pages 5-12. 1998.
10. Chalendar, G., Grau, B. : SVETLAN or How to Classify Words Using their Context. Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management table of contents. Pages 203 - 216. 2000. Springer-Verlag.
11. Cimiano, P., Hotho, A.,Staab, S. : Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. JAIR - Journal of AI Research, Volume 24: 305-339, 2005.
12. Hahn, U., Markó, K.: Ontology Evolution by Text Understanding 15th European Conference on Artificial Intelligence (ECAI'02): Workshop on Machine Learning and Natural Language Processing for Ontology Engineering.France. 2002.
13. Hahn, U., Romacker, M. : The SynDiKATe text knowledge base generator : in Proceedings of the 1st International Conference on Human Language Technology Research. San Francisco.Morgan Kaufmann, pp. 328 - 333. 2001.
14. Meadch,e, A., Staab, S. : Semi-Automatic engineering of ontologies from text. In proceeding of the 12th International Conference On Software and Knowledge Engineering. USA, 2000.
15. Bisson, G., Nedellec, C., Canamero, D.: Designing clustering methods for ontology building: The Mo'K workbench. In Proceedings of the First Workshop on ontology learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI). Germany. 2000.