Competence Mining for Collaborative Virtual Enterprise

Ali Harb[1], Kafil Hajlaoui[2], Xavier Boucher[1]

[1]Ecole des Mines de Saint Etienne
158 cours Fauriel, centre G2I, 42023 Saint Etienne, France
{harb,boucher}@emse.fr
[2]INIST - CNRS
2 allée du Parc de Brabois CS 10310
54519 Vandoeuvre les Nancy, France
Kafil.hajlaoui@inist.fr

Abstract. In a context of decision-aid to support the identification of collaborative networks, this paper focuses on extracting essential facets of firm competencies. We present an approach for enrichment of competence ontology, based on two steps where a novel effective filtering step is utilized. First we extract the correlation between terms of a learning dataset using the generation of association rules. Second we retain the relevant new concepts using an extracted semantic information. The suggested approach was tested on an ontology of mechanical industry competencies. Experiments were performed on real data, which show the usefulness of our approach

Keywords: Enterprise networks, Ontology, Data mining, Association rules, Mutual Information.

1 Introduction

Several research work deal with the formalization of characteristic data concerning potential partners for networked organizations (Camarinha-Matos and Afsarmanesh 2003, Plisson.J et al. 2007, Ermilova & Afsarmanesh 2007). Most of these approaches are adapted to a semi-closed environment defined by a Virtual Breeding Environment (VBE). The VBE provides a pre-selection of potential partners, which facilitates the share of the information required.

This paper focuses on a complementary step, which consists in providing a decision aid support for identifying potential VBE when they do not exists. This research is based on the hypothesis of an open environment of potential partners to build VBEs. Typically, this issue appears when you have to analyse a regional business area so as to identify potential collaborative networks among firms. As a consequence of the initial hypothesis, the identification of potential collaborative partners will be based on the use of public information, available through the public web sites of the companies. This assumption induces specific information extraction mechanisms.

In reference to an economic approach of firm coordination (Richardson, 1987), this research focuses the information extraction procedure on two key coordination factors: the activity fields of the companies and their internal competencies (see justification in (Hajlaoui & al, 2008a)). The information on company activities and competencies will be further used, at a second stage, in order to generate new knowledge on the potential settlement of VBEs. The overall approach of decision aid has been already described in (Hajlaoui & al, 2008b).

In the current paper, we only focus on extracting key information concerning company competencies, from public and non-structured data (company web sites). Due to the complexity of the concept of competence, a semantic oriented approach is required for the extraction. In this objective, we describe in this paper the procedure of automated enrichment of an existing competency ontology, which will be used later on for an information extraction procedure based on the use of syntaxic and semantic patterns. The use of such patterns will make possible a semantic and pragmatic treatment of available data. To reduce the complexity of ontology creation, we focus our research on enhanced semantic automatic ontology enrichment specific.

This extraction approach follows the three layer model of (fig.1, Ehrig and al, 2005):

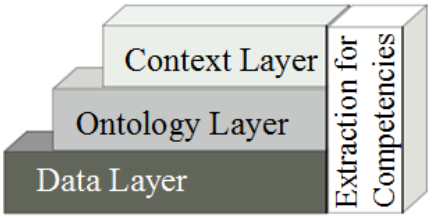

Fig 1-Three layer model for extraction

In this model the data layer structures the data source. In our procedure, data are extracted from company web sites. The html pages are cleaned and standardized with a lemmatization process. At this stage, data are considered as lexical entities. The ontology layer handles the semantic concepts to be extracted from the data sources. The concepts and their relations can be represented through an ontology which helps sharing information among distinct companies thanks to its genericity, and which make possible various semantic treatments. The third context layer is oriented towards a finer pragmatic analysis of the information to be extracted. In our approach, this layer will be developed by the formalisation and use of patterns.

It is important to note that, in this decision aid approach, there is no need to extract a precise map of company competencies. The final information expected is a similarity measure among the competency fields of distinct companies. In that objective, the ontology based extraction mechanism we will use intends to extract a set “traces” concerning the competencies of a given company.

The objective of the paper is to present a method used for automatic enrichment of competency oriented ontology. In section 2, we provide some key characteristic of the initial competency ontology that ahs been formalized and some insights on ontology building methodologies. In section 3 we describe the methodology for discovering

potential conceptual term to be integrated within the ontology. In section 4, we briefly introduce some learning concepts of this approach.

2 Ontology for Competence Mining

The identification of company competencies is a key factor for a decision support to build collaborative networks. The ontology requires rigorous conceptualization of the notions related to firm competencies. ARCHONTE ontology building methodology (Bachimont & al., 2002) was chosen, because it defines precise cognitive principles and mechanisms for each step of ontology formalization. This method makes possible a good semantic representation of the conceptual network of the ontology through a « semantic normalisation » mechanism.

From a company website, it is often impossible to extract a relevant structured and exhaustive identification of firm competencies. However it remains possible to extract a sufficient set of competence "traces", which makes possible to assess with a good degree of precision an indicia of competence similarity among firms. To implement the extraction mechanism, an ontology of "competence traces" has been formalized. This ontology is composed of two sub-ontologies: a *generic ontology* and a *domain ontology*. The former provides a very generic conceptual model of "competence traces", which is independent of any business sector. However because of this high generality level, the generic ontology is not directly applicable for information extraction. Then, it is complemented by a *domain ontology,* which constitutes an extension of the generic ontology, specific to some selected business sector (mechanical industry in our case). The "competence traces" concepts of this domain ontology are directly linked to the applied field of the mechanical industry. Furthermore, at the lower level of the domain ontology, each ontological concept is associated to specific terms of the domain, further used in the extraction mechanism as "competence trace *identifiers*". The *identifiers* will be used to detect the presence of a competence trace concept within document corpora (extracted from company websites). This detection of "competence traces" is called the activation process applied to the ontology conceptual classes.

To fulfill the objective of extraction information, the direct detection of terms in a corpus is not a sufficient condition. To confirm the presence of an ontology concept (associated to one or several *identifiers*) in the corpus, we have to cope with context-dependent linguistic phenomena which can change the semantic of the *identifiers*: synonyms, antonyms and, more generally, semantic similarity among terms make necessary to use specific semantic techniques. To deal with such ambiguity phenomena we propose the use of linguistic patterns.

UNICOMP system is the implementation of our approach which aims at extracting company competence traces, using public information available on websites.. The coupling between ontology and patterns present a real added-value for the richness of information extraction. This research has already been further developed to integrate these results in a Decision Support System for collaboration Network Building using competence similarities measures (Hajlaoui, 2009). In the next section, an automatic or semi-automatic enrichment of the ontology and the pattern base, using case-

learning techniques would be able to increase the semantic representation of competence and the overall performance of the system.

3 Discovering Potential Conceptual Term

We distinguish two types of methods for the discovery candidate's concepts: the first is based on statistical calculations using several measurements to select potential terms according to their distribution in corpora (Agirre & al., 2002), (Steinmetz & al., 2006),(Parekh & al., 2004), such as complex measurements such as mutual information, tf-idf, etc, or the use of statistical laws of term distributions (Neshatian & al., 2004). These various proposals enable identifying new candidate of enrichment terms, but do not allow placing them in ontology, without a tiresome human intervention (Jorio & al., 2007). The second type refers to syntactic methods which determine the grammatical function of a word or a word group within a sentence. It is based on the following assumption: the grammatical structure reflect semantic dependences (Bendaoud , 2006). It uses the grammatical functions of a word or a word group to present new concepts. Such methods present the drawbacks of identifying only the relations labelled by verbs. Other approaches also use syntactic patterns (Pekar & al., 2002). The extracted terms illustrate the new potential enrichment concepts.

To insert the provisional terms within an existing ontology, it is essential to detect the relations among these new terms and initial ontology concepts.), (Steinmetz & al., 2006) proposes a statistical approach based on the frequent co-occurrence of candidates terms with the concepts of initial ontology. A lack of precision is noticed among the new concepts and the ontological structure. Other approaches are based on data mining techniques (Hernandez & al., 2007). Several approaches suggest using frequent correlations among the corpora terms while using extracted association rules (Srikant & al., 1997) among potential concepts (Bendaoud , 2006). Each rule expresses the relation between two or more concepts of the field. This enrichment process requires filtering considering the large number of rules generated. Consequently human intervention is necessary to define the semantic relations discovered. Other work (Han & al., 2000), (Neshatian & al., 2004) is based on the classification methods in order to bring closer the candidates terms contained in the texts to the concepts present in ontology. The principle consists in gathering terms according to their number of occurrences within corpora (Parekh & al., 2004), using a clustering method. The disadvantage of these approaches is that they do not detect the relations among the candidates terms, i.e., unfortunately they require a human intervention for the addition of these new terms.

3.1 Approach for Discovering Concept

The purpose of this section is to present our approach. The overall process is described in figure 2. It consists of three phases. In the following sub-sections these three phases are presented in detail.

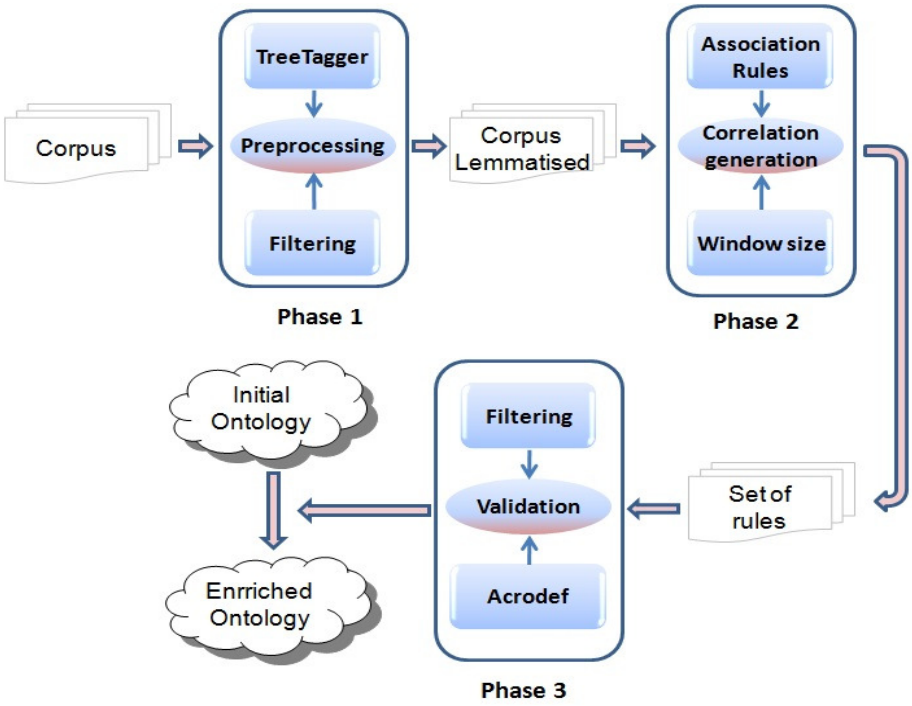

Fig 2-Approach of ontology enrichment for competence extraction

3.2 Corpora Preprocessing

To extract and to preserve the context of sentences contained in the corpora (Hajlaoui, 2009) initially obtained from web sites, we manually built a list of stop words to be filtered (e.g. , mailto, next, previous, GIF, jpg . . .). These words are always frequent in the corpus in spite of the extraction text phase (passage of the format HTML with the textual format. This list is used to filter and clean the corpora. According to the expert analysis of the field for the initial ontology concepts and the texts of the companies describing the field of ontology, the choice of the words to be kept is according to three grammatical categories (Verb, Noun and Adjectif). In order to do that, we treated morpho-syntactically the corpora using TreeTagger (Schmid, 1994). This application enables us to obtain the category and lemma of each word. A filtering step according to these three categories was made to build a new corpora based on words considered grammatically relevant.

3.3 Window Size Creation and Concept Extraction

The objective is to search within the corpora the words which are correlated to the concepts of the ontology. Perhaps, a brief explanation of our use of the term *window*

size is in order. A window size is a set of words that surround a given concept. For this, from the processed corpora, we seek correlations between ontological concepts and words of documents so as to enrich the ontology with more relevant and useful words. The question that arises at this point is: How to search for words correlated to the ontological concepts? The answer consists in two steps:

1) construction of Window size;
2) generation of association rules;

3.4 Generation of Association Rules

Nevertheless, in order to obtain more relevant concepts, consider the following hypothesis: the more a word is close to an initial concept, the more likely this word has strong semantic correlation. Thus, sentences are defined by considering window sizes (WS). A window size is a set of words that include one or more initial concept of ontology. In other words, as aforementioned a set of words that surround concepts. Our goal is to identify in sentences how to represent these WS. For instance, if WS is set to 1 that means that a sentence is composed by one word before and one after the pivot concept.

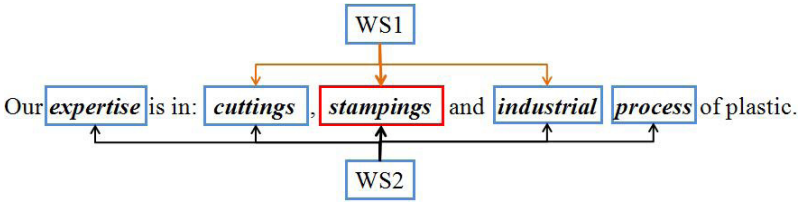

Fig 3-Contextual localisation of competence concept

In figure 3 the pivot is stampings. Using WS whose size 1, we get the following transaction "stampings, cuttings and industrial", and by specifying a WS= 2 : "stampings, cuttings, expertise, industrial and process. These windows are the transactions for the next step. The second step is the generation of the association rules. In order to detect the semantic correlation between the terms in documents and the ontology's concepts, an association rule algorithm has been adapted (Agrawal & al., 1994) to our concern. More formally, let $I = \{i_1, ..., i_n\}$ a set of terms, and D a set of sentences, where each sentence corresponds to a subset of elements of I. An association rule is thus defined as $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \phi$. The support of a rule corresponds to the percentage of sentences in D containing $X \cup Y$. The rule $X \rightarrow Y$ has a confidence ratio c, if c% of sentences from D containing X also contain Y.

4 Experimentation

The initial domain ontology contains a large number of concepts, and the enrichment is a tedious task that requires a lot of time. Consequently, in a first time we chose to test on a single conceptual domain of the initial ontology to measure the performance of our approach. The initial list of concepts described in Table I.

Table 1. Result of learning concepts.

Initial Concepts	Generated Association Rules	Learned Concepts
Assembly, Production	Assembly !- injection(3.2, 100.0)	Adjusting, Injection
Boiler, Grinding	Stampings !- cutting (4.0, 90)	Metal, Welding
Turning, Surface	Forging !- Threading (1.7, 100.0)	Metrology, Simulation
Stamping, Treatment	Area !- metrology (2.1, 100.0)	Tools, Threading
Manufacturing, Usage	Area!- welding (1.6, 100.0)	expertise
Forging, Thermal	Machining !- expertise (2.3, 100.0)	Molding
Milling, Machining	Machining !- accuracy (1.8, 100.0)	Turning
Laser, Carry	Machining !- installation (2.9, 100.0)	Cutting

As seen from, the fact of using windows where the pivots words are the initial concepts of the ontology, and the using nouns, verbs and adjectives enables to improve considerably the detection of correlation in texts. In the same way, the use of filters significantly retained correlated relevant concepts. It should be noted that our approach was sufficiently automatic to be applied in various fields and thus extract the significant concepts elsewhere. The first results obtained are promising: we discovered and placed suitably new concepts. The analysis of the ontology obtained showed that the whole of the concepts discovered is coherent since most of them could be attached to ontology via the rules obtained. These results were also validated by the expert of the field.

5 Conclusion

In this paper, we present a new approach of ontology enrichment based on data mining techniques specifically the association rules and the use of extracted semantic information to retain relevant new concepts. Our approach is based on three steps:

1) Preprocessing of a textual corpora (cleaning and lemmatising: treetagger).

2) Creating window size to promote the correlation between corpora words and concepts of the ontology, then the application of the APRIORI algorithm to extract association rules according to validated parameters (support, confidence).

3) Automatic Analysing and filtering of generated rules to keep those that are relevant to the initial domain ontology.

The obtained enrichment results seem auspicious for the chosen domain ontology. This approach is sensitive to the domain studied: it is important to have a corpora that describes the subject domain accurately, so there is a very rich language with nouns and adjectives available. Although the company websites do not represent the ideal

source for building such a corpora for the domain studied (companies competency in the mechanical industry), but we have had promising results.

References

1. Camarinha-Matos, LM and Afsarmanesh H., *Elements of a base VE infrastructure Computers in industry*, vol. 51, 139-163, 2003.
2. Ermilova E., Afsarmanesh H., *Modeling and management of profiles and competencies in VBEs*. Journal of Intelligent Manufacturing 18, 561-586, 2007.
3. Ehrig M. and Peter Haase and Nenad Stojanovic and Mark Hefke. Similarity for Ontologies -- a Comprehensive Framework. *Proceedings of the 13th European Conference on Information Systems, Regensburg, Germany, (2005)*.
4. Hajlaoui k., Boucher X., Mathieu M., Information Extraction procedure to support the constitution of Virtual Organisations. Research Challenges in Information Science, RCIS'2008, Marrakech, 2008a.
5. Hajlaoui K., Boucher X., Mathieu M., Data Mining To Discover Enterprise Networks, 9 th IFIP Working Conference on Virtual Entreprises (PRO-VE'08) Poznan, POLAND, 8-10 September 2008b.
6. Plisson, J. Ljubic P, Mozetic I, Lavrac N. (2007). An ontology for Virtual Organisation Breeding Environments. *To appear in IEEE Trans. On Systems, Man, and Cybernetics*.
7. Richardson R. J., (1987), Directorship interlocks and corporate profitability, *Administrative Science Quarterly*, vol.32, pp.367-386.
8. B. Bachimont, A. Issac, and R. Troncy. Semantic commitment for designing ontologies. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), LNAI 2473 :114–121, 2002.
9. K. Hajlaoui, thèse, Dispositifs de recherche et de traitement de l'information en vue d'une aide à la constitution de réseaux d'entreprises. EMSE Saint Etienne, France, Dec 2009.
10. E. Agirre, O. Ansa, E. Hovy, and D. Martinez, "Enriching very large ontologies using the www," Proceedings of ECAI 2000 workshop on Ontology Learning, 2002.
11. A. F. A and R. Steinmetz, "Ontology enrichment with texts from the www," Proceedings of the 2nd Semantic Web Mining Workshop at ECMLI/PKDD, WS'02, 2006.
12. V. Parekh and J. G. T. Finin, "Mining domain specific texts and glossaries to evaluate and enrich domain ontologies,"Proceedings of the International Conference of Information and Knowledge Engineering, 2004.
13. K. Neshatian and Hejazi, "Text categorization and classification in terms of multi-attribute concepts for enriching existing ontologies," Proceedings of the 2nd Workshop on Information Technology and its Disciplines, pp. 43–48, 2004.
14. L. D. Jorio, L. Abrouk, C. Fiot, D. Hrin, and M. Teisseire,"Enrichissement d'ontologie bas sur les motifs squentiels," Actes de la Plateforme AFIA 2007, Atelier Ontologies et gestion de l'htrognit smantique, 2007.
15. R. Bendaoud, "Construction et enrichissement d'une ontologie partir d'un corpus de textes," Actes des Rencontres des Jeunes Chercheurs en Recherche d'Information (RJCRI'06), Lyon, pp. 353–358, 2006.
16. A. M. V. Pekar and S. Staab, "On discovering taxonomic relations from the web," Journal of Information Retrieval, Contextual Information Retrieval Systems, vol. Springer Verlag, pp. 301–322, 2002.
17. N. Hernandez, J. Chrisment, and D. Egret, "Modeling context through domain ontologies," Journal of Information Retrieval, Contextual Information Retrieval Systems, vol. 10, pp. 143–172, 2007.
18. R. Srikant and R. Agrawal, "Mining generalized association rules," Future Generation Computer Systems, vol. 13, n. 23, pp. 161–180, 1997.
19. E. Han and G. Karypis, "Centroid based document classification : Analysis and experimental results," Proceedings of The 4th European Conference of Principles of Data Mining and Knowledge Discovery, pp. 424–431, 2000.
20. H. Schmid, "Probabilistic part-of-speech tagging using decision trees," Conference on New Methods in Language Processing, Manchester, UK., 1994.
21. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," VLDB'94, 1994.