14 DATA MINING TO DISCOVER ENTERPRISE NETWORKS

Kafil Hajlaoui, Xavier Boucher and Mihaela Mathieu
Ecole Nationale Supérieure des Mines de Saint Etienne, FRANCE, hajlaoui@emse.fr

Within the framework of Virtual Organisations (VO), a decision aid approach was developed to support the identification of collaborative corporate networks. This approach is based on an automated procedure of information extraction to identify key features of potential partners. The added value of this research is to operate in an "open universe" of potential partners, using the company internet sites as the main source of information on firms. The key features extracted concern the activity fields and the competencies of the firms.

1. INTRODUCTION

Data Mining appeared as a new discipline which complements statistics and information technology fields (Friedman, 1997)-.. Data Mining has been described as "*the nontrivial extraction of implicit, previously unknown, and potentially useful information from data*" (Frawley and al., 1992). Data mining has merged with Knowledge Discovery in Databases (KDD) (Hébrail and Lechevallier, 2003). Enterprises amass and refine immense amounts of data routinely: customer profiles, production stock, manufacturing levels, etc. Many data-processing means are implemented nowadays in order to help decision-makers deal with this information overload. Examples abound: data warehouses provide a support for decisional information systems; data mining solutions extract new knowledge from these data warehouses, etc.

The success of Small & Medium-size Enterprises (SME) confronted with the Global marketplace relies more and more on their ability to put into practice business intelligence. The deployment of business intelligence solutions turns out to be essential for many strategic decisions for instance innovation in complex products, or collaboration in or through Virtual Organizations (VO).

In this paper, we present the first results of an approach aimed at facilitating the constitution or set up of Virtual Organizations (VO). The objective is to make a direct use of public information available through company web sites, in order to broadly analyze potential co-operative opportunities. To the best of our knowledge, most previous publications in this field have concerned a semi-closed environment defined by a Virtual organization Breeding Environment (Ermilova & Afsarmanesh, 2007). Such VBEs pre-selects potential partners, who have provided pre-structured information in order to further evaluate collaborative possibilities. The added value of the proposed approach presented, is that there is no need for any pre-treatment phase. Our method is applied within a full and open environment of potential partners. Potential collaborator identification is based on the use of public information available on the web. This assumption leads to specific information extraction mechanisms.

The rest of the paper is organized as follows. In section 2 we present an overview of the whole approach. In section 3, we focus on information extraction mechanisms to identify correctly activity fields of the company. In section 4, we use these first results for a first level of decision aid based on an enterprise clustering procedure. In conclusion, a brief discussion is presented detailing some of the advantages of the suggested approach and limitations that still need to be overcome.

2. OVERVIEW OF THE APPROACH

Advances in computer networking technology constitute key success factors for the management of Virtual Organizations (Dewey and al., 1996), (Gilman and al., 1997). A typical application is the creation of VO which emerges from the members of a Virtual Breading Environment (Lavrac, 2005), (Camarinha-matos and Afsarmanesh, 2003). Such approaches often use competency analyses to help in identifying potential collaborations among companies. However "complementarities of activity sectors" constitute another partnership criterion already discussed in the literature. It provides one of the main motivations for co-operation, in addition to the traditional motivation of sharing or pooling costs (Géniaux and al., 2003). In the current research we consider both aspects: complementarities on activity sectors and similarity of competencies. As underlined by figure 1, this method works in two steps: first information extraction is employed to identify key characteristics of the company; second a decision aid phase which uses these key characteristics as input to discover potential collaboration alliances.

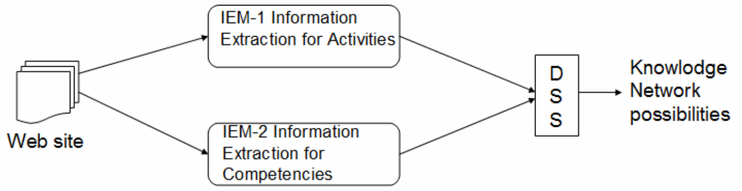

IEM-1 Information Extraction for Activities
Web site
IEM-2 Information Extraction for Competencies
D S S
Knowlodge Network possibilities

Figure 1- Two extractions mechanisms required

This approach is dedicated to an open environment of potential VO partners. The initial data comes directly from public information available on company internet sites. This is without any restrictions or preconditions imposed by or on potential candidate firms. The extraction mechanisms depicted in figure 1 focus on the 2 main characteristics we have selected to consider collaboration potentials: supplementary activities and competency similarities. We also describe in the paper (section 4) how activity and competency features can be used, in a second step, to generate new knowledge on construction of collaborative networks. Because of space requirements, this paper will only focuses on activity complementarities.

3. DATA MINING FOR ACTIVITY FIELD IDENTIFICATION

The following sections focus on extraction mechanisms applied to the identification of "enterprise sector of activity" (IEM-1 on figure 1). The overall extraction procedure is synthesized by figure 2. This paper briefly details the procedure and concentrates on demonstrating the feasibility by the results described in section 3.1 and 3.2. The main steps are Extraction-lemmatization, Indexation then Similarity matching. The reader can refer to (Hajlaoui et al., 2008) for a complete justification.

The extraction procedure uses an external semantic resource (thesaurus) built from the standardized French NAF code (Nomenclature of French Activity). The NAF is a standard frame used in France to describe the main sector of activity for all French companies. This NAF frame is presented as a hierarchical tree, referencing all potential activity fields into classes and sub-classes. The use of this reference frame facilitates the automation of the indexation algorithms. The aim is to use web site information in order to classify any given company in the NAF tree.

The extraction of information on the activities proceeds statistically, using a controlled indexing approach. As mentioned the NAF code is used as a thesaurus reflecting a semantic and conceptual representation of all potential fields of activity.

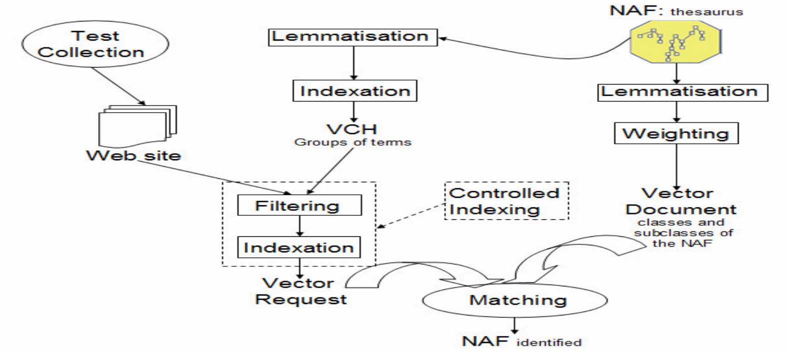


Figure 2-Structure of the information extraction procedure

First this thesaurus is used as a database for the search engine. Second, it constitutes an external semantic resource meant to improve the precision of expression for information needs. Here, an "information request" concerns only one firm's web site. Informally, this request could be expressed by "Which NAF code is correct for a given firm?". This request will be associated to a "request-vector" (figure 2), which is a set of terms extracted from the company web site. The request-vector is built by a controlled indexing with SMART (Salton's Magic Automatic Retrieval Technique)[1]. The controlling vocabulary is provided by the NAF code.

The code NAF is also used as external semantic resource. Each of the classes or subclasses of the NAF code[2] is represented by a specific document. This document is described by a vector called "document-vector". The document-vector contains a set of characteristic terms for each NAF class or subclass. The terms are associated to specific weights according to their relevance in the field.

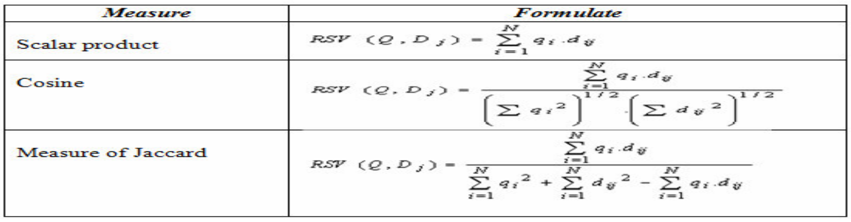
[1] ftp://ftp.cs.cornell.edu/pub/smart/, discussed in (Hajlaoui and al, 2008)

[2] The NAF code has a hierarchical structure of classes and subclasses of activity fields. Example of activity field classes : C28 class- Work of metals, C29 class- Manufacture of machines and equipment, C28.1 subclass - Manufacture of metal elements for construction, etc…

Finally the request-vector is to be matched with the various document-vectors, using a similarity measure. This matching process aims at identifying the most suitable (similar) NAF code to represent the company.

3.1 Results

The experimental phase began with a test-corpus composed of 25 companies with well-known NAF code. The objective is to measure the feasibility and of the reliability of the approach. Thus the objective is to evaluate the results by comparing the code discovered by our approach and the actual NAF code of the company. Initially, the test was conducted with three broad NAF classes in the field of the mechanical industry: classes C28 (metal work), C29 (manufacture of machines and equipments) and C34 (automotive industry). Then the experiment was repeated with more refined subclasses to verify the ability to detect more precise NAF codes.
The similarity comparison of the request and document is done in two stages. First, an attempt to find the most relevant NAF class for the company is effectuated. The highest similarity score is selected. Second, an exploration of the sub-classes is accomplished in order to recompute a more precise NAF code. Three similarity measuring functions are typically used for this comparison: scalar product, cosine, Jaccard (table 1). These 3 alternatives were tested on the corpus.



Measure	*Formulate*
Scalar product	$RSV\ (Q, D_j) = \sum_{i=1}^{N} q_i . d_{ij}$
Cosine	$RSV\ (Q, D_j) = \frac{\sum_{i=1}^{N} q_i . d_{ij}}{\left(\sum q_i^2\right)^{1/2} . \left(\sum d_{ij}^2\right)^{1/2}}$
Measure of Jaccard	$RSV\ (Q, D_j) = \frac{\sum_{i=1}^{N} q_i . d_{ij}}{\sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} d_{ij}^2 - \sum_{i=1}^{N} q_i . d_{ij}}$

Table 1: Typical similarity measures

Most of the results of this experiment have already been detailed in (Hajlaoui and al, 2008), therefore only a synthesis will be given. The first results of NAF code identification by selecting the highest similarity score are very encouraging. The general performance for this extraction mechanism is 76% (percentage of actual NAF code identified). The robustness of the system was tested by dealing with companies outside the 3 NAF classes considered (C28, C29, C34). The system proved to be accurate, with null similarity measured in such cases. However further testing remains necessary to broadly validate the results. The approach did not seek to be exhaustive since the objective was to test feasibility. Before launching more extensive experimentation which would cover all the classes of the NAF code, the performance of the extraction mechanism requires optimization, based performance indicators as developed below.

3.2 Performance of the information extraction mechanism

In the domain of information research systems, the performance evaluation of tool/method is usually based on two indicators: Recall and Precision. The recall is calculated by the number of common elements among the “relevant-documents” and the “found-documents”, divided by the number of relevant-documents. The precision is measured by the number of common elements among the relevant-documents and the found-documents, divided by the number of found-documents. Another performance indicator proposed is the frequency of precisions equal to 0. These situations must be avoided, since a null precision means no relevant-documents.

To measure the performance indicators for each request, 2 sets of relevant-documents and found-documents must be defined. Here, the definition of these sets and then the final performance of the system will depend on two factors which have been further tested:

- The similarity measure. The performance impact of each of the 3 performance measures defined above has been tested: scalar product, cosine, Jaccard. For one request (web site), the similarity provides several scores indicating the similarity of the request-vector to each of the document-vectors.
- The way to determine the "found-documents". This set can be reduced to 1 element with the highest similarity score (scoremax), but it can also gather a set of documents for which the score belongs to an interval [scoremax-α%, scoremax], with different values of α. The values 10, 20 or 33 were tested. The maximum tolerance 33% was determined referring to the possible values for the terms' weights.

Some key results are shown in figure 3. The figure underlines the 3 indicators for different configurations of similarity measures and parameter α. These experiments lead one to conclude that the best performances are obtained with α=33%, and with the function cosine. In that configuration the actual company NAF code is found in 92% of the requests on classes and 76% on subclasses.

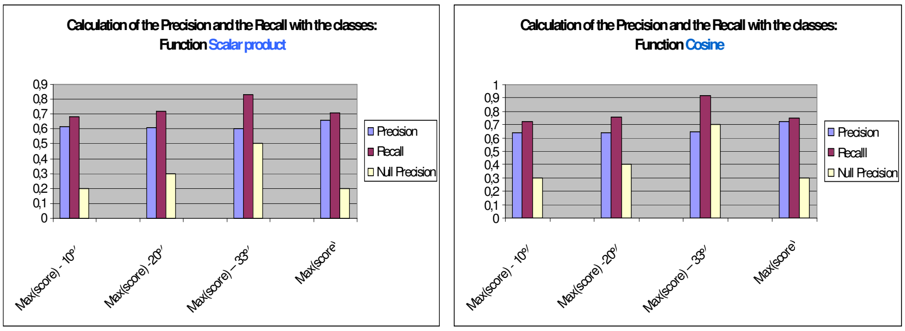


Figure 3 - Performance evaluation of the extraction system.

4. APPLICATION TO NETWORK BUILDING

The final objective of the research is to identify potential collaboration opportunities among companies. In that perspective the complementarities among industrial sectors of activity is a quite common factor when considering collaboration (Milgrom, 1997), (Frayret and al, 2003). The following section refers to the approach defined in (Burlat, Benali, 2007) which was selected in an already published state of the art (Hajlaoui et al., 2008). The authors proposed a decision-making aid based on clustering algorithms which can be applied using the results of the suggested extraction mechanisms as an input.

4.1. Method

To model if the activities are complementary graph theory is exploited to facilitate the mathematic treatment required. A graph is used to represent a set of companies and their complementarities. Each node in the graph corresponds to one company, and the arc between two nodes represents an evaluation of the degree of complementarity. Here we have referred to a definition of complementarity which considers that two sectors of activity are complementary when they can both be used to achieve integrated products/services available on the market. According to this definition, a complementary link between two companies is symmetrical (non-oriented graph). Referring to this definition, some expertise from specialists of the mechanical industrial domain is necessary to formalise an initial generic matrix of complementarity degrees among the various activity fields defined by the NAF code. This matrix is used as a generic data to evaluate complementary links among the companies of the collection.

4.2 Overview on the clustering algorithm

The clustering algorithm proceeds by a progressive elimination of the smallest weighted arcs (i.e., arcs with low degrees of complementarity). First, the weakest arcs are detected and then eliminated. Several steps occur, increasing the elimination threshold each time, thus isolating sub-groups of enterprises which should be more and more complementary. The number of steps for the algorithm will be chosen according to the number of clusters and the degree of complementarity expected (Benali and Burlat, 2004). The advantages of this algorithm are that it does not eliminate the strongly weighted arcs and is rather easy to apply.

The intent of this partition algorithm is to isolate strongly inter-connected sub-graphs based on information loss minimization (loss of arcs, loss of potential complementarity). These sub-graphs will represent set of very complementary companies: later, this information on activity fields clusters (associated with other information) will be used to justify potential collaborations. To apply the algorithm, the degree of complementarity must be determined among all the companies. The generic matrix of complementarity degrees is applied for that purpose. The 25 companies of the test collection are distributed on 8 NAF activity fields. To test the algorithm, one company was chosen to represent each of these 8 sectors. The initial complementarity graph is displayed in the figure 4.

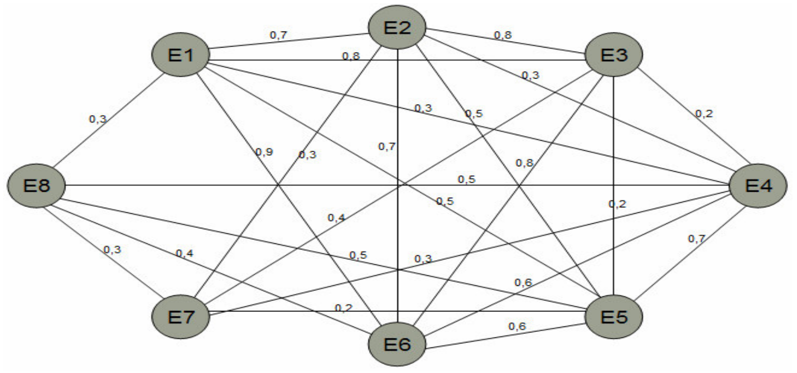


Figure 4: The case study of company graph

With this complementarity graph as a starting data, the algorithm can be applied. The following table presents the results of the various partitioning steps.

Steps	Arc(k)	Removed arc's	I	Sub-groups	Quality
1	0,1	∅	0	∅	Weak
2	0,2	{E7, E5} {E5, E3} {E3, E4}	0,05	∅	
3	0,3	{E1, E4} {E1, E8} {E2, E4} {E2, E7} {E4, E7} {E8, E7}	0,2	∅	
4	0,4	{E8, E6} {E7, E3}	0,27	{E7} {E1, E2, E3, E4, E5, E6, E8}	Average
5	0,5	{E1, E5} {E5, E2} {E8, E4} {E8, E5}	0,44	{E7} {E8} {E1, E2, E3, E4, E5, E6}	well
6	0,6	{E4, E6} {E5, E6}	0,54	{E7} {E8} {E5, E4} {E1, E2, E3, E6}	
7	0,7	{E1, E2} {E2, E6} {E5, E4}	0,72	{E7} {E8} {E5} {E4} {E 1, E2, E3, E6}	
8	0,8	{E1, E3} {E2, E3} {E3, E6}	0,92	{E7} {E8} {E5} {E4} {E 1, E6} {E3} {E2}	
9	0,9	{E1, E6}	1	{E7} {E8} {E5} {E4} {E 1} {E6} {E3} {E6} {E2}	

Table 4: Partitioning the complementary graph

"I" is an indicator of quality of information contained in the partitioned graph. It evaluates the information lost at each step of the process[3]. At the initial stage the value of *I* is set at zero (all information available). At the end of the partitioning it approaches one. This means that most of the information concerning complementarities has been lost. At the same time, the indicator "Quality" is employed to evaluate the interconnection degree among the enterprises of each sub-graph identified. Based on these indicators, one can choose at which step the procedure should stop. For example after six iterations the following company clusters are obtained with *0,54* for the indicator of information loss and with the level "well" for the quality indicator:

G_1= {E7}; G2= {E8}; G3= {E5, E4}; G4= {E1, E2, E3, E6}.

[3] The evaluator is measured by the sum of the removed arcs' weights divided by the total sum of the weight arcs from the initial graph.

5. DISCUSSION AND CONCLUSION

A contribution for the construction of virtual organizations was presented. The approach consists in two stages: the first is an automatic system of information extraction focusing on identifying enterprise sectors of activity from internet sites. Its experimental results proved encouraging. The second stage identifies clusters of companies according to complementary sectors of activity. The approach is based on a partitioning algorithm applied to a graph of complementarities among companies. However this criterion of activity complementarities is insufficient to distinguish correctly operational company networks. For instance, referring only to these criteria, the 25 companies of the test collection are only distributed on 8 distinct activity sectors, limiting diversity. Future research will be oriented on considering additional criteria. The first perspective to consider will be similarity of competencies. This requires a more complex information extraction based on advanced semantic data mining, natural language treatment and ontology. The development of a semantic-oriented analysis, based on a structured model of the "competency" concept should provide significant progress.

6. ACKNOWEDGEMENT

The authors wish to thank Chris Yukna for his help in English

7. REFERENCES

1. Burlat P. and Benali M., A *methodology to characterise co-operation links for networks of firms.* Production Planning and Control Vol.18, No 2, 156-168, 2007.
2. Benali, M. and Burlat, P., *Une démarche d'analyse de la complémentarité des activités dans un réseau d'entreprises*, in 5e Conference Francophone de Modelisation et Simulation, MOSIM'04, Nantes, France, 2004.
3. Camarinha-Matos, LM and Afsarmanesh H., *Elements of a base VE infrastructure Computers in industry*, vol. 51, 139-163, 2003.
4. Dewey and al., *The impact of NIIIP virtual enterprise technology on next generation manufacturing*. In Proceedings of conference on Agile and Intelligent Manufacturing Systems, Troy, NY, October 1-2, 1996.
5. Ermilova E., Afsarmanesh H., *Modeling and management of profiles and competencies in VBEs.* Journal of Intellignet Manufacturing 18, 561-586, 2007.
6. W. Frawley and G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213-228, Fall 1992. ISSN 0738-4602.
7. Frayret J. M., D'Amours F., D'Amours S., *Collaboration et outils collaboratifs pour la PME manufacturière*, séminaire du centre de recherche CEFRIO, Quebec, 2003.
8. Friedman, J.H., *Data Mining and statistics: what's the connection*?, 1997. http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps
9. Géniaux L., Mira-Bonnardel S., *Le réseau d'entreprises : forme d'organisation aboutie ou transitoire*, Revue Française de Gestion, vol. 29, 129-144, 2003.
10. Gilman C. and al., *Integration of design and manufacturing in a virtual enterprise using enterprise rules, intelligent agents, STEP, and workflow*. In Proceedings of SPIE vol. 3303, 160-171, 1997.
11. Hajlaoui k., Boucher X., Mathieu M., *Information Extraction procedure to support the constitution of Virtual Organisations. Research Challenges in Information Science,* RCIS'2008, Marrakech, 2008.
12. Hébrail, G., Lechevallier, Y.(2003) *Data Mining et Analyse des données in Analyse des données*, G.Govaert éditeur, Hermes, 323-355
13. Milgrom P., Roberts J., (1997) *Economie, organisation et management,* De Boeck Université, Bruxelles, Belgique.
14. Lavrac Nada and al (2005), Automated extraction and structuring of competencies from unstructured company data: Two case studies. International Conference Applied Statistics Ribno (Bled), Slovenia.