AUTOMATED STRUCTURING OF COMPANY COMPETENCIES IN VIRTUAL ORGANIZATIONS

Peter Ljubič (1), Nada Lavrač (1, 2), Joel Plisson (1), Dunja Mladenić (1), Stefan Bollhalter (3), Mitja Jermol (1)

(1) Jožef Stefan Institute, Ljubljana, Slovenia

(2) Nova Gorica Polytechnic, Nova Gorica, Slovenia

(3) Virtuelle Fabrik AG, Arbon, Switzerland

ABSTRACT

Creation of virtual organizations (VO) consists of several steps. One of the early steps is finding organizations with an appropriate expertise from a larger pool of organizations, referred to as a VO Breeding Environment (VBE). This requires a good knowledge of each company' competencies and thus emerges the need for a structured representation facilitating the choice of appropriate partners. Ontologies provide a structured representation of knowledge that is potentially useful for such a task. This paper proposes an approach to extract semi-automatically the companies' competencies from their profile data and then structure the information in the form of an ontology. This method is generally applicable for competency modeling in collaborative networked organizations. We present some existing methods and tools for ontology creation, followed by the proposed methodology to structure competencies, applied to the Bodensee industry sub-cluster of the Virtuelle Fabrik VO breeding environment.

1 INTRODUCTION

The main motivation for organizations to e-collaborate is to harness dispersed expertise, enable knowledge sharing and flexible resource management when solving a new business task. The strengths of virtual organizations (VOs) [3] lie in the range of competencies the partners are able to offer jointly through collaboration. To fully exploit this advantage, problems of efficiently storing, updating, sharing, promoting and transferring knowledge need to be solved, using appropriate knowledge management approaches. A basis for successful knowledge management is the common understanding of terms and a comprehensive representation of the knowledge. This can be achieved by the development of ontologies that can be shared among the VO partners. To introduce the context for this work, the paper first introduces the basic concepts and identify some of the existing methods and tools for ontology creation that are applicable in the context of *Virtual Organization Breeding Environments* (VBEs, [3]).

This paper proposes an approach to automated extraction of companies' competencies from their profile data, and semi-automatically structure these competencies in the form of an ontology in order to facilitate the representation of the competencies. The approach is illustrated with an experiment on the Bodensee industry sub-cluster of Virtuelle Fabrik (VF), an industrial cluster of mechanical engineering companies.

2 ONTOLOGIES: BASIC DEFINITIONS AND OVERVIEW

An ontology can support a wide range of tasks such as natural language processing, information retrieval, database modelling, knowledge representation, etc. It provides a representation of knowledge, which can be used and re-used, in order to facilitate both the comprehension and the communication between different actors. These actors can be software agents or people that need to access or share a piece of information. The most basic type of ontology is only a set of terms representing a controlled vocabulary (e.g. a glossary), which are the terms that people agree to use when dealing with a common domain. By providing definitions, an ontology helps people to use the same words for expressing themselves and thus understanding themselves more easily.

The content of an ontology depends both on the amount of information and on the degree of formality that is used to express it. Generally, we distinguish two main types of ontologies: lightweight and heavyweight [6]. A lightweight ontology is a structured representation of knowledge, which ranges from a simple enumeration of terms to a graph or taxonomy where the concepts are arranged in a hierarchy with a simple (specialization, is-a) relationship between them. A heavyweight ontology adds more meaning to this structure by providing axioms and broader descriptions of the knowledge.

Different approaches have been used for building ontologies, most of them using manual methods. An approach to building ontologies was set up in the CYC project [9], where the main step involved manual extraction of common sense knowledge from different sources. [13] propose a methodology for manual ontology construction consisting of four stages: purpose identification, ontology building, evaluation and documentation.

Most of the recent work on semi-automated ontology construction addresses the problem of extending the existing WordNet ontology using Web documents [1], using clustering for semi-automated ontology construction from parsed text corpora [2], and learning taxonomic e.g., *is-a*, [4] and non-taxonomic e.g., *has-part*, relations [10].

Our scope is limited to lightweight ontologies, where concepts are companies' competencies and instances are documents containing company' profiles.

3 ONTOLOGIES IN THE VBE CONTEXT OF VO CREATION

The purpose of VBEs is to support fast virtual organization (VO) creation: VOs are formed from a cluster of organizations in a VBE when a new business opportunity arises. As a VO is formed to meet a specific demand, it dissolves once the demand is fulfilled.

One of the difficulties in VO creation is partner selection with appropriate skills and competencies. The most important strength of the whole VO lies in the choice of partners with supplementary/complementary knowledge, skills and tools. The wider is the network of knowledge the more competitive is the VO. Therefore, the network can be highly dynamic with partners entering and leaving a VBE.

For the sake of marketing a VBE and for VO creation through appropriate partner selection, a VO broker has to have access to a knowledge repository, where the information about company resources, process costs, resource availability and company profiles in terms of skills, competencies, products and past projects are stored. To be able to manage the network and perform successful knowledge management, appropriate tools have to be selected. These include ontologies and knowledge maps. While ontologies enable appropriate domain conceptualization achieved in the consensus of involved ontology developers, knowledge maps [5] provide "a visual representation of a knowledge domain according to criteria that facilitate the location, comprehension or development of knowledge". Due to the complex and dynamic nature of VBEs, information gathering and VBE/VO analysis and modeling are best supported using advanced knowledge technologies, including data, text and web mining, decision support, as well as link and social network analysis. Web crawling is a useful means for data gathering, while visualization is of ultimate importance for the presentation of obtained results.

4 A METHODOLOGY FOR AUTOMATED EXTRACTION OF COMPANY COMPETENCIES

This section presents the proposed approach, aimed towards automated extraction of company competencies, using agglomerative hierarchical document clustering. Clustering methods aims to build clusters (groups) of objects so that similar objects fall into the same cluster (internal cohesivity) while dissimilar objects fall into separate clusters (external isolation). A class of clustering methods are *hierarchical clustering* methods [8], whose purpose of is to fuse objects (instances) into successively larger clusters, using a measure of (dis)similarity.

A typical result of this type of clustering is a hierarchical tree or *dendrogram*. A dendrogram is a binary tree where single objects form the leaves of the tree and each node of the tree represents a cluster of similar objects. The further the node is from the tree root, the more similar the items are under the node. For each node, the dissimilarity at which the respective objects were joined together into a new single cluster is called the *cluster level*. It is used to determine the most appropriate number of clusters that reflects the real structure in the data: at the point where the difference between successive cluster levels is maximal, the dendrogram is `cut', producing the partition where each cluster is the most internally cohesive and there is the highest external isolation between clusters. To implement hierarchical clustering, the (dis)similarity between objects and between clusters of objects has to be defined. The most frequent measure used in document clustering [12] is the so-called *cosine similarity* between bag-of-words vector representations of documents, where 1/0 represents the presence/absence of a word in the document.

The proposed approach to semi-automated ontology construction consists of the following steps:

1. pre-process the data to get a list of words representation of documents, further transformed into a bag-of-words vector representation of documents
2. apply agglomerative hierarchical clustering to construct a dendrogram in which each cluster is represented by the most characteristic words (using the *TFIDF* word quality measure known from information retrieval and text mining)
3. select the `best' set of clusters (either according to expert's background knowledge, or automatically by cutting the dendrogram at the point where the difference between successive cluster levels is maximal), and represent the leaves of the dendrogram by the most characteristic words
4. visualize the output of clustering, together with the corresponding word descriptions
5. manually form an ontology from the obtained cluster hierarchy by labeling the clusters by appropriate concept names, and their hierarchical dependencies by appropriate relationships between concepts (e.g., *part-of*, *subset-of*, …)

5 AN EXPERIMENT USING THE VIRTUELLE FABRIK INDUSTRY CLUSTER DATA

Company profiles of a 20-partner VFEB Bodensee industry sub-cluster of the Virtuelle Fabrik virtual organization breeding environment was made available for the experiment. Each company is described by its name, number of employees, products, services and their core competencies.

5.1 Data pre-processing

Text mining approaches usually require a pre-processing phase in which the document representation format is changed from the free text form to the bag-of-words representation (commonly used in text processing), possibly preceded by stop word elimination and stemming. In our experiment, this pre-processing step was simplified by simple elimination of stop words and some manual

elimination of meaningless adjectives, resulting in a list of words. Each company was also assigned a unique numeric identifier (see Table 1). Note that the company identifier and company name were not used as input information for text clustering used in the experiment.

1 AE&P	2 ALWO	3 Bachli	4 Bruggli	5 Beni
6 Buchler	7 Ccb	8 Flube	9 KBB	10 Heese
11 Innotool	12 Knobel	13 IPG	14 M+S	15 OMB
16 Pantec	17 Schar	18 SMA	19 Sulzer	20 Wiftech

Table 1: *Company identifiers assigned to company names.*

During pre-processing, the input data was transformed into word lists, representing simplified descriptions of the original text describing the companies. Selected company descriptions are shown below.

AE&P Entwicklung Konstruktion Bereich Maschinen Anlagenbau Lieferung Komplettanlagen Konstruktions einsatze direkt Kunden Projektmanagementmandate Gesamtlosungen Automation Entwicklung Handgeraten Breites CADKnowhow Autocad Bravo Catia Euklid

ALWO Zulieferfirma Halbleiterindustrie Werkzeugbau Sonderanlagen Baugruppenmontage Serien Stuck Kleinteilefertigung drehen frasen Montage Prufprotokoll

Bachli Transformatoren Drosseln Speisegerate Flexibilitat Schnelligkeit Sicherheit Normenerfullung Produktanforderungen EN UL CSA Normen

Bruggli Druckerei Informatik Internetdienste Fahrradanhanger Techn Textilprodukte Gurten Taschen Planen Industrie Kleingeratemontage Mechanische Bearbeitungen Offsetdruck Informatik Internetdienste Techn Textilfertigung Mechanik Montage Profil Rohrbiegen Frasen Bohren Baugruppenmontage

Beni …

5.2 Experimental results

The dendrogram of Figure 1 was induced from 20 VFEB company profiles (text documents) by using the agglomerative hierarchical clustering method available as part gCLUTO, a publicly available interactive clustering, visualization and analysis system [11]. It was built automatically, together with the lists of most representative words representing each document/cluster, whereas company identifiers (in curled brackets) were added manually, in order to simplify the interpretation of the obtained structure.

The dendrogram of Figure 2 was produced from the dendrogram of Figure 1, by cutting the dendrogram at the level where differences between successive cluster levels are maximal. This resulted in six company clusters, described by automatically extracted keywords describing the clusters. Curled brackets list original documents/companies belonging to a cluster (e.g., cluster 29 includes descriptions of companies 10, 13 and 14), while cluster numbers [0]-[5] corresponding to clusters as marked in cluster visualization in Figure 3.

```
38                                  [herstellung, entwicklung, management]
|-----35                             [entwicklung, management, engineering]
|    |-----------29                 *** [management, engineering, sql]
|    |          |---------------------7  {13} [management, modell, architektur]
|    |          |---------24          [sql, engineering, visual]
|    |                    |-----------13 {14} [web, projekte, engineering]
|    |                    |-----------8  {10} [visual, sql, wonderware]
|    |-------31                       *** [entwicklung, produkte, technischer]
|            |-------27               [produktentwicklung, unterstutzung, ...]
|            |      |-----------------1   {7} [produkte, technischer, bestehender]
|            |      |-----------------12 {16} [motion, schnelle, problemlosungen]
|            |---------------------20        [anlagenbau, konstruktion, maschinen]
|                             |---18  {1} [entwicklung, konstruktions, direkt]
|                             |---19 {17} [werkzeugmaschinen, unterhalt, roboter]
|-37                                  [herstellung, montage, zulieferfirma]
 |-36                                 [montage, fertigen, apparatebau]
 | |-----33                           *** [fertigen, apparatebau, komponenten]
 | |    |---------------25            [fertigen, zylinder, maschinen]
 | |    |            |-------------2  {19} [kerntechnologien, prufen, okosysteme]
 | |    |            |-------------17 {12} [zylinder, maschinen, komponenten]
 | |    |-------------------23        [metallwarenfabrikation, stahl, ...]
 | |                    |---------9 {9}[hydraulikleitungen, industriebranchen, ...]
 | |                    |---------4   {5} [stahl, metallwarenfabrikation, ...]
 | |-------32                         *** [montage, baugruppenmontage, ...]
 |        |-----------26              [montage, anlagen, automation]
 |        |         |---------------15  {6} [internationale, druckguss, fertig]
 |        |         |---------------11 {15} [montage, montagearbeiten, pumpen]
 |        |-------------------22      [baugruppenmontage, techn, ...]
 |                      |-------10  {2} [prufprotokoll, sonderanlagen, serien]
 |                      |-------5   {4} [internetdienste, techn, informatik]
 |-----34                             [herstellung, durchmesser, zulieferfirma]
      |-------30                      *** [herstellung, flexibilitat, schnelligkeit]
      |      |-----------------------0   {3} [transformatoren, sicherheit, ...]
      |      |----------------21      [herstellung, produktentwicklungen, ...]
      |                    |-----16 {11} [herstellung, konstruktion, materialien]
      |                    |-----14 {20} [produktentwicklungen, spanabhebende]
      |-----------28                  *** [durchmesser, grossteilebearbeitung, ...]
                |-------------------6   {8} [durchmesser, colletages, ...]
                |-------------------3  {18} [ausstattung, rohrbogen, ringen]
```

Figure 1: *Dendrogram of VFEB companies, starting with 20 clusters (one for each company), and merged into larger higher-level clusters. Extracted keyword descriptors of companies and company identifiers (in {}) are shown in italics. Bold cluster descriptions (*** clusters 29, 31, 33, 32, 30, 28) reappear in Figure 2.*

```
38
|-----35
|    |-----------29 *** {10,13,14}  [3] [management, engineering, sql]
|    |-------31     *** {1,7,16,17} [1] [entwicklung, produkte, technischer]
|-37
 |-36
 | |-----33        *** {5,9,12,19} [0] [fertigen, apparatebau, komponenten]
 | |-------32      *** {2,4,6,15}  [2] [montage, baugruppenmontage,
internetdienste]
 |-----34
      |-------30   *** {3,11,20}   [5] [herstellung, flexibilitat, schnelligkeit]
      |-----------28 *** {8,18}    [4] [durchmesser, grossteilebearbeitung,
plasmaschneiden]
```

Figure 2: *Dendrogram of VFEB companies obtained by agglomerating bottom level clusters into six higher-leve clustersl.*

5.3 Visualization of six profile categories

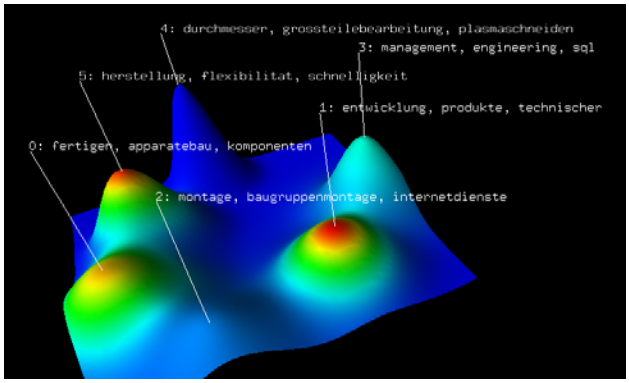

Figure 3: *Mountain visualization of six clusters, described with most descriptive words.*

The gCLUTO system [11] offers advanced cluster visualization tools, which we have used to visualize the results of VFEB clustering. The results are visualized in Figure 3. Peaks in Figure 3 represent individual clusters. The shape of each peak is a Gaussian curve, used as a rough estimate of the distribution of the data within each cluster. The height of each peak is proportional to the cluster's internal similarity. The volume of a peak is proportional to the number of elements contained within the cluster. The color of the top of a peak is proportional to the cluster's internal deviation (red indicates low deviation, whereas blue indicated high deviation). The resulting Gaussian curves are added together to form the terrain of the Mountain Visualization of gCLUTO.

5.4 Interpretation and result evaluation

Figure 3 gives a representative overview of the clusters of competencies, their strength and their homogenity. Interesting company clusters have been identified, and named by the Virtuelle Fabrik manager as follows. Cluster [1]: *Engineering, project management*; Cluster [2]: *Product, component assembly*; Cluster [3]: *Software engineering, project management*; Clusters [0], [4], [5]: *Parts manufacturing.* While clusters [1], [2] and [3] are independent clusters, clusters [0], [4] and [5] are all touching the same competencies. Cluster [4] could be identified as a sub-cluster, named *Heavy parts manufacturing.*

The results could be optimized by carefully reviewing the extracted keywords (e.g., *durchmesser* and *technischer*) which are not informative; *Flexibilität* and *Schnelligkeit* describe qualities rather than competencies). After removing non-informative words from the word list, a novel run of clustering could be performed, possibly leading to better results.

6 CONCLUSIONS AND ACKNOWLEDGEMENTS

For a complete validation of the potential of the proposed approach for the Virtuelle Fabrik industry cluster, the analyses should be extended to a wider/full set of member data. The combination of different data sources analyzed with clustering of VO creation tool, facilitating the identification of market opportunities.

Acknowledgements.
This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under ECOLEAD (IST-1506958-IP), SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP) and PASCAL Network of Excellence (IST-2002-506778). This publication only reflects the authors' views.

References

1. Agirre, E., Ansa, O., Hovy, E., Martínez, D. (2000). Enriching very large ontologies using the WWW. In Proc. of the First Workshop on Ontology Learning OL-2000, at the 14th European Conference on Artificial Intelligence ECAI-2000.
2. Bisson, G, Nédellec, C., Cañamero, D. (2000). Designing clustering methods for ontology building: The Mo'K workbench. In Proc. of the First Workshop on Ontology Learning, OL-2000, at The 14th European Conference on Artificial Intelligence ECAI-2000.
3. Camarinha-Matos, LM and Afsarmanesh H. (2003). Elements of a base VE infrastructure. J. Computers in Industry, Vol. 51, No. 2, 139-163.
4. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S. (2004). Learning Taxonomic Relations from Heterogeneous Evidence. In Proceedings of ECAI 2004 Workshop on Ontology Learning and Population.
5. Eppler, M.J. (1999). Knowledge Management Terminology. Guide, St. Gallen; mcm institute, 1999.
6. Fernandez-Lopez M., Corcho O. Ontological Engineering. Springer-Verlag, 2004
7. Gomez- Lenat, D. and Guha, R. (1990). Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley Publishing.
8. Hartigan, J.A. (1975). Cluster Algorithms. New York: Wiley.
9. Lenat, D. and Guha, R. (1990). Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley Publishing.
10. Maedche, A. and Staab, S. (2001). Discovering conceptual relations from text. In Proc. of ECAI'2000, pages 321-325.
11. Rasmussen, M. and Karypis, G. (2004). gCLUTO – An Interactive Clustering, Visualization, and Analysis System. University of Minnesota, Department of Computer Science and Engineering, CSE/UMN Technical Report: TR# 04–021.
12. Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining, 109–110.
13. Uschold, M., King, M,. Moralee, S. and Zorgios, Y. (1998). The enterprise ontology. The Knowledge Engineering Review, Vol. 13, Special Issue on Putting Ontologies to Use (eds. Mike Uschold and Austin Tate).