USING MACHINE LEARNING TO STRUCTURE THE EXPERTISE OF COMPANIES: ANALYSIS OF THE YAHOO! BUSINESS DATA

Joel Plisson (1), Dunja Mladenić (1), Peter Ljubič (1),
Nada Lavrač (1, 2), Marko Grobelnik (1)

(1) Jožef Stefan Institute, Ljubljana, Slovenia
(2) Nova Gorica Polytechnic, Nova Gorica, Slovenia

ABSTRACT

Organizations have to collaborate in order to achieve business goals which require to use a variety of domain-specific knowledge. Selection of partners with an appropriate expertise is one of the crucial tasks in the creation of virtual organizations. Partner selection can be facilitated by structuring partners' competencies in an ontology. An alternative to manual ontology construction is to group and describe similar companies according to their domain of expertise. This paper proposes an approach to automatically construct a structured representation of companies' expertise applied on the Yahoo! business data.

1 INTRODUCTION

In order to form a Virtual Organization (VO) out of companies forming a Virtual Organization Breading Environment (VBE, [3]), it is important to know the competencies of VBE partners. When the number of partners in a VBE is reasonably small, this can be handled by a human having a good knowledge about the companies. When dealing with many organizations, it becomes difficult to remember the competencies of all the partners, and it becomes necessary to model their knowledge in a form that is easily understandable and that captures the essential information.

Ontologies are a productive way to represent knowledge about a domain. They can be used to represent part of a knowledge base of a VBE that contains the knowledge shared by VBE partners. Such a representation can be achieved by identifying organizations with similar competencies and organizing them according to their domain of expertise. An example of such structure can be found in the Yahoo! business section on the Web (see http://biz.yahoo.com). In the Yahoo! business ontology, companies are grouped together into categories representing different sectors and industries.

In this paper we propose a machine learning approach to build semi-automatically a similar structure directly from company descriptions. This may result in a comprehensive representation of the market, segmented into different market categories. The presented approach may enable the structuring of company information, and potentially facilitate the selection of partners in the process of virtual organization creation.

This paper simulates a scenario of virtual organization creation from VBE partners. In this scenario, all the companies in the Yahoo! business section are treated as VBE partners, and the goal is to try to automatically structure partners' competencies from company descriptions, without considering their human labeled Yahoo! categorization. The success of automated competency structuring, using clustering tools developed in the statistical and machine learning research communities, is evaluated by comparing the results of automated clustering with the original human-labeled categories.

The structure of this paper is as follows: Section 2 presents selected approaches to ontology construction, followed by Section 3, where our approach to semi-automated ontology construction is outlined. Section 4 presents the Yahoo! business domain and the results of the experiments. We conclude with a discussion and some ideas for future work.

2 APPROACHES TO ONTOLOGY CONSTRUCTION

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.

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 [5]. 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 knowledge. The scope of this paper is limited to simple lightweight ontologies.

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 [7], where the main step involved manual extraction

of common sense knowledge from different sources. [12] 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, 10], and learning taxonomic e.g., *is-a*, [4] and non-taxonomic e.g., *has-part*, relations [8].

3 AN APPROACH TO BUILD AN ONTOLOGY OF COMPANY' COMPETENCIES

In this paper we use machine learning techniques, namely hierarchical clustering of text documents, to structure the expertise of companies based on short company descriptions available on the Web in the Yahoo! business section. The approach to build an ontology of company' competencies, proposed in this paper, is based on document clustering [11] and cluster visualization, followed by the most important part of ontology construction (step 5) involving human intervention.

The proposed procedure consists of the following steps:

1. pre-process the data to get the `bag-of-words' representation of documents
2. initialize the first cluster to the whole document set
3. apply hierarchical *k*-means clustering as follows:
   - for each cluster do
   - if a stopping criterion is satisfied, stop splitting the cluster and describe the cluster with the most characteristic words
   - else repeat step 3 on the documents belonging to this cluster
4. visualize the output of clustering
5. manually form an ontology from the obtained hierarchy of clusters through detecting inconsistencies, possibly leading to manual re-engineering of clusters, naming the clusters by appropriate concept names, and their hierarchical dependencies by appropriate relationships between concepts (e.g., *part-of*, *subset-of*, …), and interpreting and evaluating of the results of clustering.

4 EXPERIMENTS IN STRUCTURING THE EXPERTISE OF COMPANIES

We have partially implemented the proposed machine learning approach, described in Section 3, through the use of two document clustering systems, both performing hierarchical k-means clustering [11] and providing the visualization of the generated clusters. In this way, we have implemented only steps 1 to 4 of the procedure outlined in Section 3. As there was no expert involved in the experiments, we were unable to implement step 5. As an alternative, we were only able to evaluate the results of clustering by comparing the results of clustering to the existing human-labeled Yahoo! ontology, an evaluation approach which is obviously unrealistic in real-life expertise modeling scenarios.

4.1 The experimental domain: Yahoo! business data

We have performed the analysis of Yahoo! business data, which we have downloaded from the Yahoo! business sector (see http://biz.yahoo.com). The experimental data set consists of descriptions of 7107 companies (brief summaries of companies' competencies). The length of the summaries varies from 180 to 1031 characters, averaging in approx. 842 characters per description. In Yahoo!, companies are structured into 12 *sectors*, which are further divided into 102 *industries*. For example, the *Healthcare* sector is divided into four industries: *Biotechnology & Drugs, Healthcare Facilities, Major Drugs, Medical Equipment & Supplies*. The number of industries in each sector and the distribution of companies over the sectors are shown in Table 1.

Sector	Industry	Industries	Companies
Basic Materials	Gold&Silver, Iron&Steel, …	11	429
Capital Goods	Aerospace & Defense, …	7	361
Conglomerates	Conglomerates	1	29
Consumer Cyclical	Footwear, Tires, …	12	318
Consumer Non-Cyclical	Beverages, Crops, …	8	232
Energy	Coal, Oil & Gas, …	4	310
Financial	Insurance, S&Ls/Savings, …	10	1212
Healthcare	Facilities, Major Drugs, …	4	860
Services	Advertising, Restaurants, …	25	1486
Technology	Hardware, Software, …	11	1578
Transportation	Airline, Railroads, …	6	150
Utilities	Electric, Water, …	3	142
Total		**102**	**7107**

Table 1: *Names of Yahoo! sectors, some industries, the numbers of industries (per sector) and companies (per sector).*

4.2 Experimental goals, the selected approaches and results

Trying to build an ontology of 7107 company summaries manually, we would have faced the problem of not knowing the characteristics of different business areas (e.g. banking, software, healthcare etc.), which would have disabled us of producing a relevant structuring of the domain. In this experiment, our goal was thus to automatically construct a hierarchical structuring of companies into distinct categories, with the potential (in step 5 of the approach presented in Section 3) to be interpreted as an ontology of Yahoo! companies.

We applied document clustering to automatically build a hierarchy of sets of documents, i.e., a hierarchy of company groups with a subset-of relationships between the groups of companies. In hierarchical k-means clustering, used in our approach, all companies are split into k groups; each group is further split into k subgroups, based on the similarity between company descriptions. In our experiments we used two different clustering and visualization systems.

The first system [6] provides a two dimensional visual representation of document groups generated by k-means hierarchical clustering. The system performed several levels of 2-means clustering, and the stopping criterion (minimum number of companies in the clusters) was set to 1000. This resulted in a company hierarchy of 5 levels containing 11 nodes as shown in Figure 1. The main idea of tiling visualization is to split the rectangular area, representing all the companies, into sub-areas according to the size (numbers of instances) of sub-clusters. When a stopping criterion is satisfied, keywords describing the clusters are assigned to the leaves. The levels of the hierarchy are denoted by the ellipses connecting similar groups.

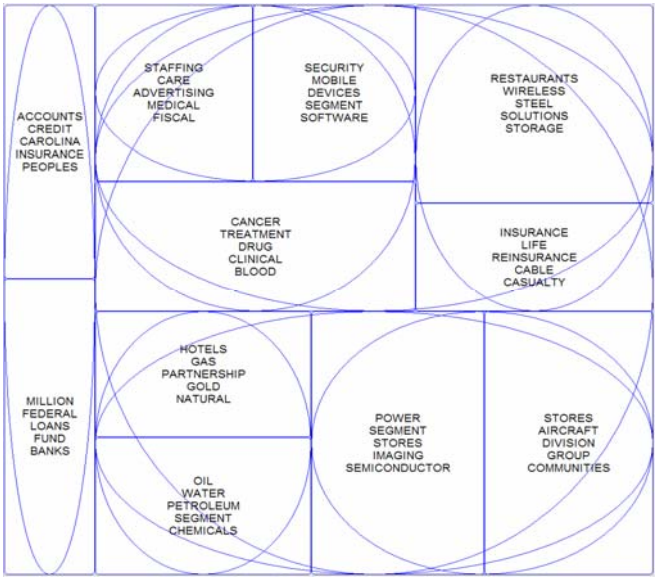

Figure 1: *Tiling visualization of companies' competencies, where the companies are organized in several hierarchical levels.*

The second system, gCLUTO [9], first performs stop-words removal and stemming in text pre-processing, followed by k-means clustering, using a predefined number of clusters of leaf-level nodes as the stopping criterion. In real-life scenarios, appropriate setting of the stopping criterion is a non-trivial task. In our experiment we have selected k equal to 12, the number of Yahoo! sectors, as one of our goals was to reconstruct the available Yahoo! business sector ontology. In gCLUTO's mountain visualization (shown in Figure 2), each peak represents an individual cluster: peak height is proportional to cluster's internal similarity, grayscale tone is proportional to cluster's internal deviation (darker tones indicate lower deviation), and peak volume is proportional to the number of elements in the cluster.

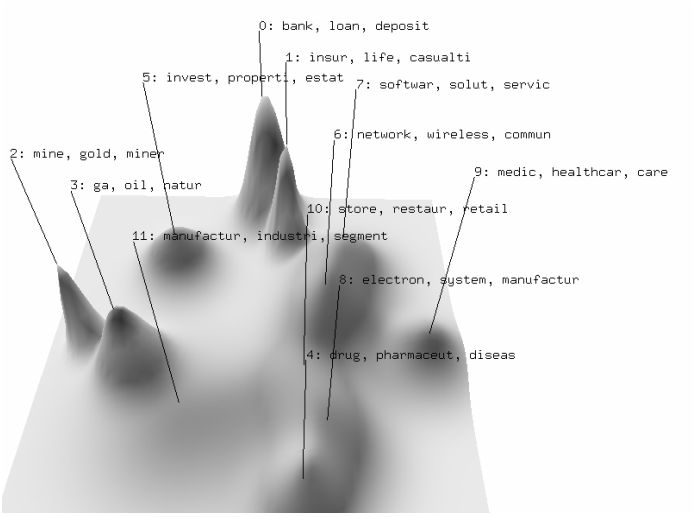

Figure 2: *Mountain visualization of 12 top-level clusters with inter-cluster similarity (ISim) represented by the peak heights*.

4.3 Evaluation of results

Without expert assistance we were unable to implement step 5 of the proposed semi-automated ontology construction methodology. Instead of intuitively naming the clusters by sector/industry names, we have - to the best of our capacity - manually aligned clusters to Yahoo! sectors, by comparing Yahoo! sector and industry names to the cluster keywords, and have evaluated the success of clustering on the scale 1-5, based on the number of keywords which – in our opinion – describe the sector.

While the first system resulted in a relatively weak correspondence between clusters and the Yahoo! sectors/industries, the cluster keywords proposed by the second system were pertinent enough to define distinct clusters that can be relatively easily understood and interpreted. Therefore we have concentrated on the results of the second approach by further analyzing the distribution (Table 3) of companies over the Yahoo! sectors in each cluster. To do so, the companies were labeled with their respective sector, and then the distribution of labels in each cluster was examined.

We can notice that clusters with higher inter-cluster similarity (ISim) contain more companies with the same label. In some cases, companies are spread among two or more different sectors. For instance the companies of cluster 6 (described by keywords *network*, *wireless*, *communications, internet, service*) are spread over sectors *Technology* and *Services*, which are closely related.

Id	ISim	Healthcare	Technology	Services	Basic Mat.	Financial	Cons. Cyc.	Capital Goods	Cons. Non-C.	Utilities	Transport	Energy	Conglom.
0	0.190	1	6	19	2	765	0	3	1	0	0	1	1
1	0.174	1	2	7	0	184	1	6	0	0	0	1	2
2	0.151	0	3	10	108	0	0	5	0	0	0	11	0
3	0.097	1	7	12	12	17	3	26	3	122	24	277	1
5	0.089	1	6	211	7	150	1	14	1	0	2	4	1
4	0.068	447	36	8	10	2	1	4	3	0	0	0	0
6	0.063	4	267	370	1	15	5	10	0	0	2	0	0
7	0.060	7	590	212	4	33	5	12	4	0	9	1	1
9	0.052	348	48	40	4	17	0	1	6	0	1	0	1
8	0.053	6	541	49	27	3	54	71	3	0	1	1	10
10	0.035	24	11	446	10	9	131	18	151	0	1	1	1
11	0.030	20	61	102	244	17	117	191	60	20	110	13	11

Table 3: *Results of System 2 - the distribution of 12 clusters among 12 sectors.*

5. CONCLUSIONS AND ACKNOWLEDGEMENTS

We have presented an approach to structuring the expertise of companies into a simple ontology, aimed at modeling competencies/expertise of companies from textual data. Two different systems were used to implement the proposed methodology, and two different visualization tools based on hierarchical k-means clustering of documents were applied in the experiments. To evaluate the results, we compared the results with the existing two-level Yahoo! ontology of companies. In terms of visualization, the advantage of the first system is that cluster hierarchy, represented by ellipses, is visualized in addition to leaf-level clusters. On the other hand, the mountain visualization of the second approach is especially appealing, as peak heights are proportional to cluster's internal similarity, and different color intensity is proportional to cluster's internal deviation, both being very important for estimating the success of clustering. The second system also resulted in much more cohesive clusters in terms of keywords used to describe the clusters of companies.

Despite the fact that the results are non-representative for a real-life situation in which pre-defined categories do not exist, the results of this experiment are interesting as they provide keywords representing company expertise as novel information over the human-defined Yahoo! sector categories. The results could be further improved by splitting the obtained clusters into more sub-clusters, thus achieving a complete hierarchy of companies' competencies. In addition the use of natural language processing methods could be used to provide additional information for word sense disambiguation, leading to improved clustering results and improved keyword extraction.

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. Elements of a base VE infrastructure. J. Computers in Industry, Vol. 51, No. 2, 139-163, 2003.
4. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S. (2004). Learning Taxonomic Relations from Heterogeneous Evidence. In Proc. of ECAI 2004 Workshop on Ontology Learning and Population.
5. Fernandez-Lopez M., Corcho O. Ontological Engineering. Springer-Verlag, 2004
6. Grobelnik, M. and Mladenić, D. (2002). Efficient visualization of large text corpora. In Proc. of the seventh TELRI seminar. Dubrovnik, Croatia
7. Lenat, D. and Guha, R. (1990). Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley Publishing.
8. Maedche, A. and Staab, S. (2001). Discovering conceptual relations from text. In Proc. of ECAI'2000, pages 321-325.
9. 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.
10. Reinberger, M-L. and Spyns, P. (2004). Discovering knowledge in texts for the learning of DOGMA-inspired ontologies. In Proc. of ECAI 2004 Workshop on Ontology Learning and Population.
11. Steinbach, M., Karypis, G., Kumar, V. (2000). A comparison of document clustering techniques. In Proc. of KDD Workshop on Text Mining, pp. 109–110.
12. 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).