Metodološki zvezki, Vol. 3, No. 2, 2006, 369-380

Automated Structuring of Company Profiles

Peter Ljubič[1], Nada Lavrač[2], Dunja Mladenić[1], Joel Plisson[1], and Igor Mozetič[1]

Abstract

Selection of partners with appropriate competencies, resources and skills is one of the crucial tasks in the creation of virtual organizations. Partner selection can be facilitated by structuring competencies in an ontology which provides a shared conceptualization. Manual ontology construction is a time and resource consuming activity. Alternatively, there are text mining, conceptual clustering and visualization tools available that can be used for semi-automated ontology creation. This paper proposes a methodology and presents tools which facilitate competency structuring from unstructured company data. These tools have been applied to the reconstruction of the Yahoo! business ontology.

1 Introduction

In order to form a *Virtual Organization* (VO) out of companies that participate in a cluster of organizations which are willing to collaborate – called a *Virtual Organization Breading Environment* (VBE) (Camarinha-Matos and Afsarmanesh 2003) - 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 manually by a knowledgeable VO broker. However, when dealing with many organizations, it gets difficult to be aware of the competencies of all the partners, and it becomes necessary to model their competencies in a form that is easily understandable, can be shared, and that captures essential partners' profile information.

For the sake of VBE marketing and for VO creation through appropriate partner selection, the 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 successfully manage the knowledge network,

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia; {peter.ljubic, dunja.mladenic, joel.plisson, igor.mozetic}@ijs.si

[2] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia; and University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia; nada.lavrac@ijs.si

appropriate tools have to be selected. These include domain specific ontologies and knowledge maps.

An *ontology* (Gruber, 1993) enables appropriate domain conceptualization achieved through the consensus of involved ontology developers. It can be used to represent a part of a knowledge base which is shared by VBE partners. Such a representation can be constructed by identifying organizations with similar competencies and organizing them according to their domains of expertise. An example of such structure is the Yahoo! business ontology available on the Web. In the Yahoo! business ontology, companies are grouped together into categories representing different sectors and industries.

While ontologies are a productive way to represent knowledge about a domain, *knowledge maps* (Eppler, 1999) also provide a useful “visual representation of a knowledge domain according to criteria that facilitate the location, comprehension or development of knowledge". The process used to gather the information needed for knowledge map construction - called *knowledge mapping* - can use as its input the information available in the constructed ontologies, the information gathered in the VBE knowledge repository, as well as the information about the business environment gathered from outside of the VBE. 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 has high utility for the presentation of obtained results to the human expert.

Manual ontology construction is a time and resource consuming activity. Alternatively, there are text mining, conceptual clustering and visualization tools available that can be used for semi-automated ontology creation (Bisson et al., 2000; Cimiano et al., 2004; Grobelnik and Mladenić, 2005; Reinberger and Spyns, 2004). This paper proposes a methodology and presents several tools which facilitate semi-automated competency structuring from unstructured company data. These tools have been applied to the reconstruction of the Yahoo! business ontology.

The structure of this paper is as follows: Section 2 gives the motivation and proposes a five steps methodology of semi-automated ontology creation. In Section 3, the reconstruction of the Yahoo! business ontology is presented. We conclude with a discussion and some ideas for future work.

2 Motivation and methodology

A proper approach to ontology creation, which conceptualizes a domain of discourse, requires careful knowledge engineering. An ontology uses a common vocabulary and structures the knowledge in classes and subclasses, including relevant properties and relations between objects. Descriptions of individual companies then correspond to individual instances in the ontology. Once the base ontology is agreed upon and created centrally, individual companies can insert their relevant data independently.

2.1 Motivation

Both stages of ontology creation are demanding in terms of human resources. For instance, the creation of the top-level ontology of the CyC project (Lenat and Guga 1999) took years. The Yahoo! business ontology is much simpler, and was accordingly easier to create, but is considered of moderate quality by some experts. It is even more optimistic to expect individual companies, e.g., SMEs, to carefully and extensively describe their competencies and skills in terms of the common ontology vocabulary.

These limitations of human engineering resources motivate the need for the development of semi-automatic tools for ontology creation. One should take advantage of the existing information already available on the Web and extract relevant facts about the companies. Obvious sources are home pages, but additionally, legal registers and business associations' public data can be used. It is clear that the quality of the Web data is of varying quality and can not be compared to the manually crafted descriptions. However, the processes of focused Web crawling, data extraction and structuring can be automated, thus relieving valuable human resources.

2.2 Methodology

The proposed methodology for semi-automated ontology construction consists of the following steps:

1. **Data gathering** (yields textual data).
   a. Data can be gathered *manually* through questionnaires filled-in by companies.
   b. Alternatively, data is also available on the Web, including company home pages and public registers. In this case, a data gathering method employed is *focused Web crawling* (Ester et al., 2001).
2. **Preprocessing** (of textual data into the bag-of-words representation).
   Raw textual data is processed as follows:

   a. Markup tags and stop-words are eliminated.
   b. Stemming or lemmatization. Each word is presented in the "normal" form by its lemma or stem, e.g., by eliminating suffixes and prefixes (Porter 1980).
   c. Transformation into the bag-of-words (BOW) representation where a document is encoded as a feature vector with word frequencies as elements.

Elements of vectors are weighted with IDF weights (Deerwester et al. 1990). All the i-th elements are multiplied with IDFi = log(N/DFi), where N is the total number of documents and DFi is document frequency of the i-th word (the number of documents in which the i-th word appears). Such vectors are also called TFIDF vectors.

3. **Structuring** (of bag-of-words into clusters).
   Structuring of the BOW representations is performed by document clustering (Steinbach et al., 2000). We applied document clustering to automatically build a hierarchy of companies, based on their descriptions, with a subset-of relationships between the groups of companies. In our experiments we used two different k-means hierarchical clustering systems: TextGarden implementation of hierarchical clustering (Grobelnik and Mladenić, 2002) and gCLUTO (Rasmussen and Karypis, 2004). In the hierarchical k-means clustering, all companies are split into k groups; each group is further split into subgroups, based on the similarity between company descriptions. The result of clustering is a taxonomic ontology, which is a simple tree structure with classes, subclasses, instances and their properties.
4. **Visualization** (of taxonomic ontology).
   Many methods were developed for the visualization of text documents or high dimensional data in general. Some examples are Themeview, Themeriver, Topic Islands (http://www.pnl.gov/infoviz), and Self-Organizing maps (http://websom.hut.fi/websom/). In this work we applied two visualization methods: tiling visualization (Grobelnik and Mladenić, 2002) and mountain visualization (Rasmussen and Karypis, 2004).
5. **Evaluation and elaboration.**
   Result evaluation by domain experts and – if available - comparison to existing ontologies.

3 Automated reconstruction of the Yahoo! business ontology

The goal of this case study was to evaluate the utility of the proposed methodology and of the available knowledge engineering tools for ontology consrtruction. To this end, we have automatically reconstructed the Yahoo! business ontology and compared it to the original, manually created one.

We have partially implemented the proposed methodology of semi-automated topic ontology construction, described in Section 2, through the use of two document clustering systems, both performing hierarchical k-means clustering and visualization of the generated clusters. In this way, we implemented steps 1 to 4 of the procedure outlined in Section 2. We evaluated the results (step 5) by comparing the automatically generated clusters with the existing human-labeled

Yahoo! ontology, thus estimating the success of the semi-automatic reconstruction of the Yahoo! ontology from unlabeled textual company descriptions.

The specific steps taken in terms of the proposed methodology were the following:

1. Data were gathered from the Yahoo! http://biz.yahoo.com.
2. Textual descriptions of companies were transformed into the standard bag-of-words document representation.
3. Structuring was performed by the application of two clustering algorithms (one as implemented in TextGarden and the other as implemented in gCLUTO) which both yielded simple taxonomic ontologies.
4. Visualization was done in the form of tiling and mountain visualization.
5. Results were evaluated by the comparison to the original Yahoo! taxonomy.

3.1 Yahoo! business data

We have performed the analysis of Yahoo! business data, extracted from the Yahoo! business sector on the Web (http://biz.yahoo.com). The extracted data set consists of textual 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 manually 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.

3.2 Clustering and visualization of the results

The goal of the experiment was to automatically reconstruct the manually constructed Yahoo! ontology from unlabeled textual company descriptions (ignoring sector and industry labels), and evaluate the success by comparing the automatically generated hierarchical structure with the original human generated ontology. The experiment was thus aimed at verifying whether - instead of manually building an ontology of 7107 company summaries from scratch - one can automatically structure companies into distinct categories, which could be further manually elaborated into a high-quality ontology.

We applied two document clustering and visualization systems to automatically build and visualize the generated document hierarchy, i.e., the hierarchy of company groups with a “subset-of” relationships between the clustered groups of companies.

Table 1: The Yahoo! sectors, industries, the number of industries (per sector) and companies (per sector).

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**

Hierarchical k-means clustering algorithms work as follows:

```
1. initialize the first cluster to the whole document set
2. apply hierarchical k-means clustering for each cluster:
     i.  if a stopping criterion is satisfied, stop splitting the cluster and
         describe the cluster with the most characteristic words
     ii. else repeat step 2 on the documents belonging to this cluster
```

The TextGarden implementation of hierarchical clustering (Grobelnik and Mladenić, 2002) provides also a two dimensional visual representation of document groups generated by the hierarchical clustering. In the experiment, 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, visualized by tiling the space of company descriptions. The main idea of tiling visualization is to split the rectangular area, representing the companies, into sub-areas according to the size (number of instances) of sub-clusters. When a stopping criterion is satisfied, keywords describing the clusters are assigned to the leaves of the hierarchical structure. The levels of the hierarchy are denoted by the ellipses connecting similar groups.

The second system, gCLUTO (Rasmussen and Karypis, 2004), 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 the experiment we have selected k equal to 12 (the number of Yahoo! sectors) as our goal was to reconstruct the available Yahoo! business sector ontology. In the gCLUTO's mountain visualization (shown in Figure 2), each peak represents an individual cluster: the peak height is proportional to cluster's inter-similarity (ISim), the grayscale tone is proportional to cluster's internal deviation (darker tones indicate lower deviation), and the peak volume is proportional to the number of instances in the cluster.

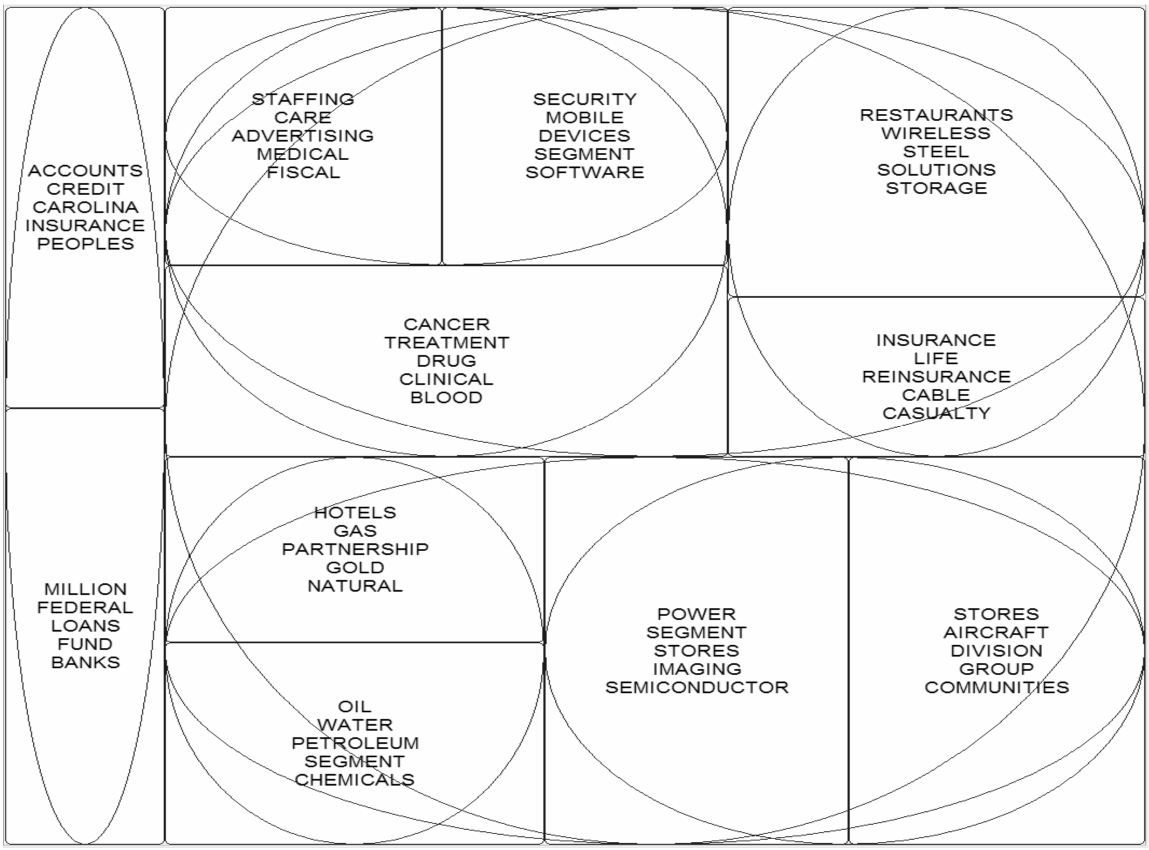

Figure 1: Tiling visualization of the Yahoo! company competencies, where companies are clustered in several hierarchical levels.

3.3 Evaluation of the results

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 automatically assigned cluster keywords. We have evaluated the success of clustering on the scale 1 to 5, based on the number of keywords which – in our opinion – describe the sector. The result of the evaluation is shown in Table 2.

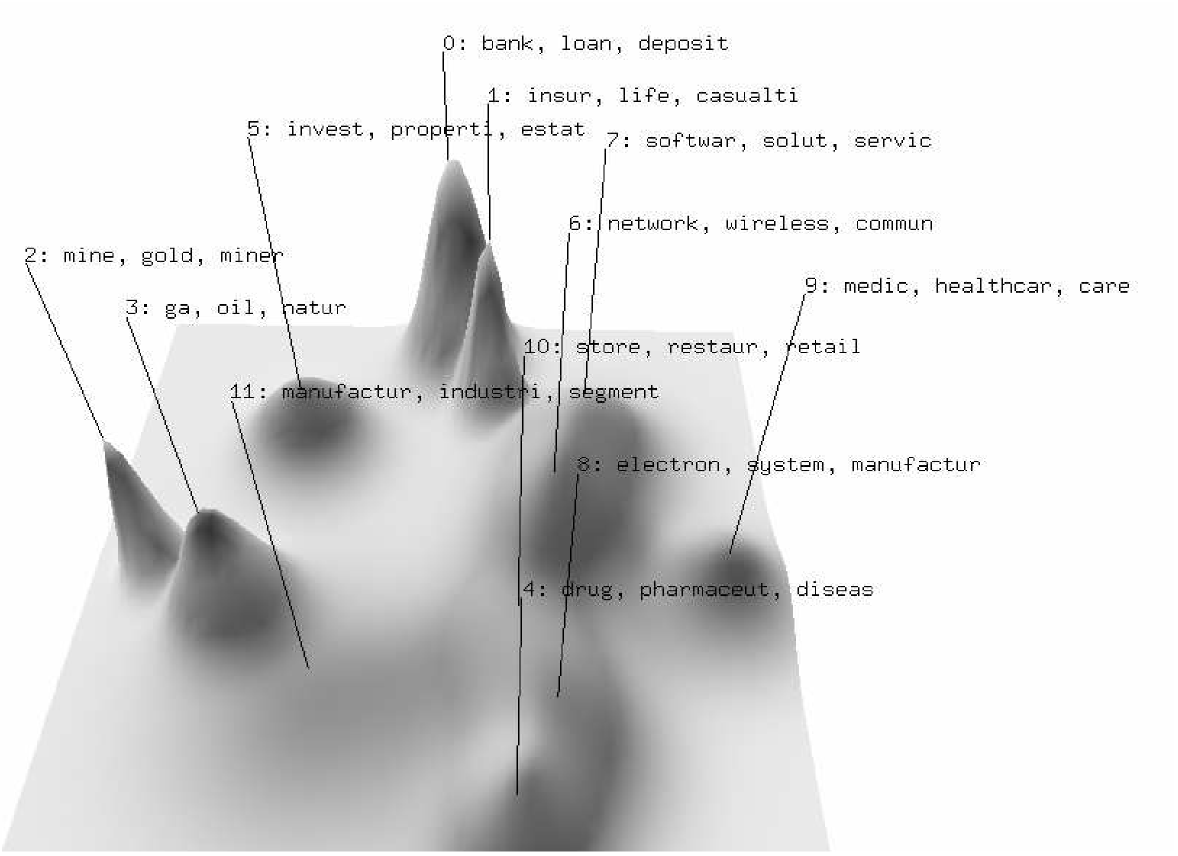
0: bank, loan, deposit
1: insur, life, casualti
5: invest, properti, estat
7: softwar, solut, servic
6: network, wireless, commun
2: mine, gold, miner
9: medic, healthcar, care
3: ga, oil, natur
10: store, restaur, retail
11: manufactur, industri, segment
8: electron, system, manufactur
4: drug, pharmaceut, diseas

Figure 2: Mountain visualization of 12 top-level clusters where the inter-cluster similarity is represented by the heights of the peaks.

The application of Text Garden implementation of hierarchical k-means clustering resulted in a relatively weak correspondence between clusters and the Yahoo! sectors/industries (evaluated by the average score 2.9). On the other hand, the cluster keywords proposed by gCLUTO (the average score 4.3) were pertinent enough to define distinct clusters that can be relatively easily understood and interpreted. Therefore, we have concentrated on the results of gCLUTO by further analyzing the distribution of companies over the Yahoo! sectors in each cluster. The companies were labeled with their respective sector, and the distribution of labels in each cluster was examined. The distribution is shown in Table 3.

The analysis of Table 3 indicates 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.

Table 2: Clusters generated by the two clustering systems (each cluster is described by keywords and evaluated by a score) mapped to Yahoo! sectors and industries.

Yahoo! sectors and industries	**Text Garden clusters** **Keywords (Score)**	**gCluto clusters** **Keywords (Score)**
Basic Materials Gold&Silver, Iron&Steel, …		• mine, gold, miner, exploring, property (4) • manufacturing, industry, segment, product, steel (1)
Capital Goods		
Conglomerates		
Consumer Cyclical		
Consumer Non-Cyclical		
Energy Coal, Oil & Gas, ...	• hotels, gas, partnership, gold, natural (2) • oil, water, petroleum, segment, chemicals (3)	• gas, oil, natural, energy, exploring (4)
Financial Insurance, S&Ls/Savings, …	• accounts, credit, Carolina, insurance, people (3) • million, federal, loans, fund, banks (5) • insurance, life, reinsurance, cable, casualty (4)	• bank, loan, deposit, mortgage, finance (5) • insurance, life, casualty, reinsurance, property (5) • invest, property, estate, real, trust (4)
Healthcare Facilities, Major Drugs, …	• cancer, treatment, drug, clinical, blood (5) • staffing, care, advertising, medical, fiscal (2)	• drug, pharmaceutical, disease, treatment, cancer (5) • medic, healthcare, care, health, hospital (5)
Services Advertising, Restaurants, …	• restaurants, wireless, steel, solutions, storage (1)	• store, restaurant, retail, brand, food (5)
Technology Hardware, Software, …	• security, mobile, devices, segment, software (4) • power, segment, stores, imaging, semiconductor (2)	• network, wireless, communication, internet, service (5) • software, solution, service, information, management (4) • electron, system, manufacturing, semiconductor, equipment (5)
Transportation Airline, Railroads, …	• stores, aircraft, division, group, communities (1)	
Average score	2.9	4.3

4 Conclusions and future work

We have presented a methodology to structure the expertise of companies into a simple competency ontology from textual company descriptions. Textual data is first represented using the standard bag-of-words representation. Two clustering algorithms were applied and resulting structures presented by two different visualization tools. The methodology was tested on a business data case study.

In the case study, the results were compared with the existing two-level Yahoo! ontology of companies. In terms of visualization, the advantage of the tiling visualization is that cluster hierarchy, represented by ellipses, is visualized in addition to the leaf-level clusters. On the other hand, the mountain visualization of gCLUTO is especially appealing, as the 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 gCLUTO clustering also resulted in more cohesive clusters in terms of keywords used to describe the clusters of companies.

Table 3: Results of gCLUTO - the distribution of 12 clusters among 12 sectors.

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

Despite the fact that the study does not represent 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.

Acknowledgments

We are grateful to Marko Grobelnik and Rayid Ghani for the discussion in which we chose Yahoo! ontology restructuring to be the topic of this research. The funding of the Slovenian Research Agency programme Knowledge Technologies (2004-2008) and the IST Programme of the European Community projects European Collaborative Networked Organizations Leadership Initiative

(ECOLEAD, IST-1-506958-IP), Semantically Enabled Knowledge Technologies (SEKT, IST-1-506826-IP), and PASCAL Network of Excellence (IST-2002-506778) are acknowledged. This publication only reflects the authors' views.

References

[1] Bisson, G., Nédellec, C., and Cañamero, D. (2000): Designing clustering methods for ontology building: The Mo'K workbench. In *Proceedings of the First Workshop on Ontology Learning OL-2000*, at the 14th European Conference on Artificial Intelligence ECAI-2000, 13-19.

[2] Camarinha-Matos, L.M. and Afsarmanesh H. (2003): Elements of a base VE infrastructure. *Journal of Computers in Industry*, **51**, 139-163.

[3] Cimiano, P., Pivk, A., Schmidt-Thieme, L., and Staab, S. (2004): Learning Taxonomic Relations from Heterogeneous Evidence. In *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*, 25-30.

[4] Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., and Harshman, R.A. (1990): Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**, 391–407.

[5] Eppler, M.J. (1999): Knowledge Management Terminology. Guide, St. Gallen; mcm institute, 1999, http://www.informationobjects.ch/NetAcademy/naservice/publications.nsf/all_pk/1617

[6] Ester, M., Gross, M., and Kriegel, H-P. (2001): Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies.

[7] Gruber, T.R. (1993): A translation approach to portable ontologies. *Knowledge Acquisition*, **5**, 199–220.

[8] Grobelnik, M. and Mladenić, D. (2002): Efficient visualization of large text corpora. In *Proceedings of the Seventh TELRI Seminar*. Dubrovnik, Croatia

[9] Grobelnik, M., and Mladenić, D. (2005): Automated knowledge discovery in advanced knowledge management, *Journal of knowledge management*, **9**, 132-149.

[10] Lenat, D. and Guha, R. (1990): *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Publishing.

[11] Porter, M. (1980): An algorithm for suffix stripping. *Program*, **14**, 130–137.

[12] Rasmussen, M. and Karypis, G. (2004): gCLUTO – An Interactive Clustering, Visualization, and Analysis System. University of Minnesota, Dept. of Computer Science and Engineering, CSE/UMN Tech. Report TR04–021. http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/gCLUTO.pdf

[13] Reinberger, M-L. and Spyns, P. (2004): Discovering knowledge in texts for the learning of DOGMA-inspired ontologies. In *Proceedings of ECAI 2004 Workshop on Ontology Learning and Population*, 19-24.

[14] Steinbach, M., Karypis, G., and Kumar, V. (2000): A comparison of document clustering techniques. In *Proceedings of KDD Workshop on Text Mining*, 109–110.