Unsupervised Ontology Acquisition from Plain Texts: the *OntoGain* System

Euthymios Drymonas[1], Kalliopi Zervanou[2**], and Euripides G.M. Petrakis[1]

[1] Intelligent Systems Laboratory, Electronic and Computer Engineering Dept., Technical University of Crete (TUC), Chania, Crete, Greece
{max,petrakis}@intelligence.tuc.gr
http://www.intelligence.tuc.gr
[2] Tilburg centre for Creative Computing (TiCC), University of Tilburg, The Netherlands
k.zervanou@uvt.nl

Abstract. We propose *OntoGain*, a system for unsupervised ontology acquisition from unstructured text which relies on multi-word term extraction. For the acquisition of taxonomic relations, we exploit inherent multi-word terms' lexical information in a comparative implementation of agglomerative hierarchical clustering and formal concept analysis methods. For the detection of non-taxonomic relations, we comparatively investigate in *OntoGain* an association rules based algorithm and a probabilistic algorithm. The *OntoGain* system allows for transformation of the derived ontology into standard OWL statements. *OntoGain* results are compared to both hand-crafted ontologies, as well as to a state-of-the art system, in two different domains: the medical and computer science domains.

Key words: ontology acquisition, formal concept analysis, term extraction, term similarity, term clustering, association rules, multi-word terms, OWL

1 Introduction

In modern computer science, *ontologies* are formal representations of knowledge resources in terms of *concepts* and respective *relations* which describe a certain domain. Ontology development is a time and cost consuming task, requiring specialists from several fields [1]. This high development effort constitutes, in turn, an inhibiting factor in building large scale intelligent systems. Current research in the field of automatic, or semi-automatic ontology acquisition and development aims at providing methods and solutions to this problem.

A branch of current approaches to ontology acquisition relies on text analysis techniques originating from the field of information extraction, such as the extraction of named entities and relationships, based on supervised [2, 3], or manually developed semantic analysis patterns [4, 5]. These methods involve human

** This work was carried-out while the author was with TUC

intervention and effort at various degrees, either in the development of training data or analysis resources, but they achieve relatively good representations of the domain concepts and relations. Other methods approach the ontology acquisition problem by unsupervised techniques, based on statistics and basic linguistic tools [6–8], or unsupervised information extraction [9, 10]. The principal challenge in such approaches lies in achieving a satisfactory coverage of the domain in terms of concepts and concept relations, while reducing human effort to the absolute minimum.

Within this latter framework of approaches, we present *OntoGain*, a platform for unsupervised ontology acquisition from text. In *OntoGain*, contrary to other similar approaches [6, 11, 12, 8], initial ontology concept acquisition relies on a domain independent method for multi-word term extraction [13]. Multi-word terms constitute the majority of terminological expressions and lexicalise domain concepts such that their semantics cannot be fully inferred from their single-word constituents [14]. Furthermore, multi-word terms inherently contain classificatory information, expressed as modifiers. For these reasons, their identification is expected to provide better conceptual coverage for a given domain and contribute to the subsequent relation acquisition tasks. For the acquisition of taxonomic and non-taxonomic relationships, we comparatively investigate four approaches: agglomerative hierarchical clustering and formal concept analysis, for the taxonomy development, and association rules and conditional probabilities, for the detection of non-taxonomic relations. These methods are adapted for multi-word term concept input and comparatively assessed in two different domains, the medical and the computer science domains.

Issues relating to ontology construction and *OntoGain* resources are discussed first in Sec. 2. The method is discussed in detail in Sec. 3. Evaluation results are presented in Sec. 4 followed by conclusions and issues for further research in Sec. 5.

2 Ontology Construction Methodology

The steps in ontology development may be viewed as a layer stack, where lower layers represent the basic tasks upon which rely the more complex, higher layers [12, 15, 16]. Thus, one cannot define concept relations before defining concepts.

In this perspective, *term extraction* is a basic layer for unsupervised concept acquisition from text. This task aims at identifying the set of terms which are characteristic for the domain and which will form the ontology lexicon. In our approach, terms are considered multi-word compounds. Multi-word terms constitute the large majority of term expressions. Moreover, they are vested with more compact and distinctive semantics (e.g., *"right ventricular infarction"* specifies in detail a concept different from other general mentions of myocardial infarction), and they present the advantage of lexically revealing their semantic content classificatory information by means of modifiers. For example, the compound term *"right ventricular infarction"* denotes a type of *"ventricular infarction"*, which in turn is a type of *"infarction"*. The exploitation of multi-word term informa-

tion in *OntoGain* allows for improved domain concept coverage by capturing full term phrases and by retaining, in concept representations, the inherent, in multi-word terms, classificatory information. Moreover, multi-word terms being more compact representations of the domain concepts, compared to long lists of single-word terms, allow for more efficient system performance in the subsequent steps of ontology acquisition and make the processing of large document collections faster, without exhausting system resources.

The subsequent layer, the *concept hierarchy*, constitutes the "backbone" of the ontology. This task aims at organising concepts into a hierarchical structure, a taxonomy, where each concept is related to its respective broader and narrower concepts. In *OntoGain*, we implement and comparatively evaluate two methods for unsupervised taxonomic relation acquisition: agglomerative hierarchical clustering and formal concept analysis. In our implementation of the hierarchical clustering approach, we exploit multi-word term lexicalised classificatory information to determine term similarity in clustering.

Concepts are also characterised by attributes and relations to other concepts in the hierarchy. This is the "relation" layer, where *non-taxonomic* relations are defined. These relationships are typically expressed by a verb relating pair of concepts [17]. For this layer in *OntoGain*, we again implement and compare two approaches: association rules and a probabilistic approach.

OntoGain exhibits the following two important advantages: (a) produces a semantically rich ontology of multi-word domain concepts, rather than an ontology of mere single-word terms and, (b) produces an ontology in standard OWL representation[3]. Moreover, it allows for experimentation and comparative assessment of four different approaches to the unsupervised acquisition of taxonomic and non-taxonomic relations.

3 The *OntoGain* modules

The principal modules of *OntoGain* are:

1. *preprocessing* which performs the linguistic analysis tasks required by subsequent modules,
2. *concept extraction* which identifies multi-word term phrases denoting domain concepts,
3. *taxonomy construction* which hierarchically structures the discovered concepts, and
4. *non-taxonomic relation acquisition* which enriches the taxonomy with domain specific concept relationships.

In what follows, we present these modules' implementations in more detail.

[3] Implementing the *Jena Semantic Web Framework*: http://jena.sourceforge.net

3.1 Preprocessing

For this module, we use the OpenNLP[4] suite of tools for tokenisation, POS tagging and shallow parsing. For the acquisition of word lemma information, we used the WordNet Java Library (JWNL[5]). POS tagging and lemma information are required in the concept extraction phase, whereas shallow parsing information is used in the non-taxonomic relation acquisition phase for the detection of verbal dependencies.

3.2 Concept Extraction

For its concept extraction module, *OntoGain* implements C/NC-value [13], a domain-independent method for the extraction of multi-word and nested terms.

In this approach, noun phrases are initially selected by linguistic filtering. The subsequent statistical component defines the candidate noun phrase termhood by two measures: C-value and NC-value. The first measure, the C-value, is based on the hypothesis that multi-word terms tend to consist of other terms (nested in the compound term). For example, the terms *"coronary artery"* and *"artery disease"* are nested within the term *"coronary artery disease"*. Thus, C-value is defined as the relation of the cumulative frequency of occurrence of a word sequence in the text, with the frequency of occurrence of this sequence as part of larger proposed terms in the same text. The second measure, the NC-value, is based on the hypothesis that terms tend to appear in specific context and often co-occur with other terms. Thus, NC-value refines C-value by assigning additional weights to candidate terms which tend to co-occur with specific context words.

3.3 Taxonomy Construction

Hierarchical Clustering: Hierarchical agglomerative clustering proceeds bottom-up. Initially, each term phrase is considered a cluster and, at each step, the similarity between all pairs of clusters is computed and the most similar pair is merged. The algorithm typically continues until a single cluster is formed at the top of the hierarchy. We used the group average method to compute the similarity between two clusters. In particular, the group average method computes the average similarity across all pairs of concepts within the two clusters (C_i, C_j) that will be merged:

$$sim(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} sim(x, y)}{|C_i| * |C_j|} \quad (1)$$

where x is a concept in cluster C_i and y in cluster C_j respectively. For the computation of term similarity among multi-word terms, we use the *lexical similarity* measure [18] which takes into consideration multi-word term constituents (head/modifier) and is computed according to a Dice-like coefficient formula.

[4] http://opennlp.sourceforge.net
[5] http://jwordnet.sourceforge.net

Thus, the lexical similarity sim_{lex}, between term concept x and term concept y (the heads of which are denoted by x_h and y_h respectively, and their set of constituents by C) is computed as:

$$sim_{lex}(x,y) = \frac{|C(x_h) \cap C(y_h)|}{|C(x_h)| + |C(y_h)|} + \frac{|C(x) \cap C(y)|}{|C(x)| + |C(y)|} \quad (2)$$

where the numerators denote the number of shared constituents and the denominators the sum of all constituents.

In our implementation of agglomerative clustering for the *taxonomy construction* module, the clustering process is terminated before reaching a single, top cluster. More specifically, clustering repeats as long as the merged clusters share common term heads. Furthermore, the lexical similarity measure gives credit to the shared heads between two similar multi-word terms. For this reason the created clusters consist of terms with shared heads. This cluster characteristic is exploited by *OntoGain* in appropriately labeling the top clusters of the derived concept hierarchy.

Formal Concept Analysis (FCA): FCA [19] is a popular approach for building concept hierarchies [7, 6]. It relies on the idea that objects (i.e., concepts) are associated with their attributes (i.e., characteristics). FCA takes as input a matrix specifying a set of *formal objects* and *formal attributes*. In *OntoGain*, objects are the extracted multi-word terms, whereas attributes are the associated verbs, as identified in the syntactic dependencies analysis of the shallow parser. These dependencies are used to form the *formal contexts* matrix which constitutes the input to the FCA algorithm. An example of a formal context matrix is illustrated in Table 1. Dependencies are denoted by asterisks.

Table 1. Computer Science knowledge as a formal context

	submit	test	describe	print	compute	search
html form	⋆			⋆		⋆
hierarchical clustering					⋆	⋆
text retrieval						⋆
root node		⋆	⋆		⋆	⋆
single cluster			⋆		⋆	⋆
web page				⋆		⋆

For the subsequent selection of the optimal set of concept discriminative attributes, *OntoGain* implements conditional probability measures. In a comparative study by Cimiano [6], conditional probability is reported to outperform other measures, such as pointwise mutual information (PMI) [20], and selectional strength [21].

In our experiments with conditional probability threshold values we have found that the object-attribute dependency pairs above threshold $t = 0.003$ are the optimal set of dependencies for both of our application corpora domains.

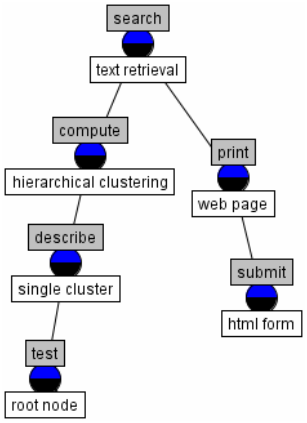

Fig. 1. Sample FCA taxonomy (Computer Science corpus)

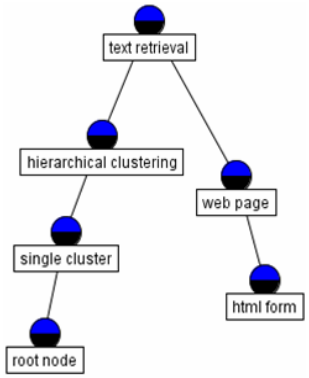

Fig. 2. Final FCA taxonomy (without attribute labels)

For the implementation of FCA, we used the Lindig's *colibri* Java library[6] which implements the Next-Closure algorithm [22]. Table 1 illustrates an example of the input lattice to FCA, whereas a tree illustration of the OWL output taxonomy is shown in Fig. 1 and 2.

3.4 Non-taxonomic Relation Acquisition

Association Rules: One of the two methods implemented in *OntoGain* for non-taxonomic relation acquisition relies on the generalised association rules algorithm, as extended for determining associations at the appropriate level of generalisation with respect to a given taxonomy [23]. In our implementation, the "subject-verb-object" dependencies set relating multi-word term concepts is enhanced with more general terms (super concepts) of the concepts it contains, based on our built taxonomy. The resulting association rules are subsequently filtered, so as to eliminate rules $X \Rightarrow Y$, where Y consists of some element in X and the rules also subsumed by $X' \Rightarrow Y'$, where X' and Y' are super concepts of X and Y respectively.

Association rules in *OntoGain* are implemented using the *predictive apriori* algorithm implementation of the Weka platform[7]. In this application, each rule is evaluated by *predictive accuracy*, a trade-off between support and confidence for

[6] http://www.st.cs.uni-saarland.de/~lindig
[7] http://www.cs.waikato.ac.nz/ml/weka

each rule maximising correct predictions [24]. In our experiments with predictive accuracy threshold values, we have found that the rules above threshold $t = 0.2$ are the best for both of our application corpora domains. Table 2 illustrates sample non-taxonomic relationships output based on association rules, for the medical corpus.

Table 2. Non-taxonomic relationships based on association rules

Domain	*Range*	*Label*
chiasmal syndrome	pituitary disproportion	cause by
medial collateral ligament	surgical treatment	need
blood transfusion	antibiotic prophylaxis	result
lipid peroxidation	cardiopulmonary bypass	lead to
prostate specific antigen	prostatectomy	follow
chronic fatigue syndrome	cardiac function	yield
right ventricular infarction	radionuclide ventriculography	analyze by
creatinine clearance	arteriovenous hemofiltration	achieve
sudden cardiac death	tachyarrhythmias	cause
cardioplegic solution	superoxide dismutase	give
bacterial translocation	antibiotic prophylaxis	decrease
accurate diagnosis	clinical suspicion	depend
ultrasound examination	clinical suspicion	give
total body oxygen consumption	epidural analgesia	attenuate by
coronary arteriography	physician assistant	perform by

Probabilistic algorithm: In this approach, reported by Cimiano et al. [25], the selection of the most appropriate non-taxonomic relationships from the set of all dependency relations found in text relies on conditional probability measures. According to this approach, first, the frequency of a dependency relation is estimated and then this frequency is propagated through the respective super concepts in a given taxonomy. The conditional probability measure is subsequently estimated as:

$$P(c|v) = \frac{f(c,v)}{f(v)} \tag{3}$$

where c is the concept, v the relation as lexicalised by a verb, and $f(c,v)$, $f(v)$ the cumulative frequencies of the concept and the relation respectively. If two or more concepts share the same conditional probability, the dependency relation of the most specific concept is selected. In our approach in *OntoGain*, dependency relations are provided by the shallow parser and concepts are multi-word terms.

4 Evaluation

In the *OntoGain* evaluation, we attempt to assess how well the extracted ontology reflects the application domains and how our *OntoGain* results compare to similar state-of-the-art approaches and hand-crafted ontologies.

We experimented with two different domain corpora, a medical and a computer science corpus. Our medical corpus was the OHSUMED[8] collection [26], containing 348,566 references from MEDLINE[9]. The computer science corpus consisted of computer science papers and articles [27].

The approach we followed for the quality assessment of the domains' representation consisted of two stages. In the first, the resulting domain ontologies are decomposed into their constituent parts (i.e., concept terms, taxonomic & non-taxonomic relations) [16] and domain experts assess each individual constituent result. In the second stage, *OntoGain* results are compared to hand-crafted ontologies. Finally, for the comparative assessment of *OntoGain* against a similar system, we applied the Text2Onto system[10] to our domain corpora.

4.1 Evaluation of ontology constituent parts

For assessing *OntoGain* in terms of *precision*, we selected the top 200 multi-word terms extracted by the concept extraction module and we proceeded to taxonomic and non-taxonomic relation extraction by all four respective methods. The domain experts then indicated the correctly acquired terms and relations. For our estimation of *recall*, domain experts examined the first 500 lines of each corpus and extracted the multi-word concepts and relations (taxonomic and non-taxonomic). These hand-crafted ontologies were compared to the respective results obtained by *OntoGain*.

The evaluation results illustrated in Table 3 show that C/NC-value performs very well in the task of identifying concepts for the ontology lexicon and in accordance with the results reported in the literature. In the taxonomy construction task, clustering clearly outperforms FCA in both corpora. Finally, association rules deliver better non-taxonomic relations when compared to probabilistic measures.

FCA results in both domains were particularly low. This was primarily due to a large number of spurious verb dependency results. Most concepts appeared with many different verbs, resulting in the formation of huge concept lattices. In FCA, concepts sharing a verb attribute are credited, so as to form taxonomic relations. However, in our application corpora, several potentially related concepts were assigned different verb attributes which resulted in either no establishment of a candidate taxonomic relation, or in the formation of erroneous and meaningless relations.

[8] http://ir.ohsu.edu/ohsumed/ohsumed.html
[9] http://www.nlm.nih.gov/databases/databases_medline.html
[10] http://ontoware.org/projects/text2onto

Table 3. Evaluation results for the Computer Science & OHSUMED corpora

Method	**Computer Science corpus**		**Medical corpus**	
	Precision	Recall	Precision	Recall
Concept Extraction				
C/NC-Value	86.67 %	89.6 %	89.7 %	91.4 %
Taxonomy Construction				
Formal Concept Analysis	44.2 %	48.6 %	47.1 %	41.6 %
Agglomerative Clustering	71.33 %	62.7 %	71.2 %	67.3 %
Non-Taxonomic Relation Acquisition				
Association Rules	72.85 %	61.7 %	71.8 %	67.7 %
Probabilistic algorithm	61.67 %	49.4 %	62.7 %	55.9 %

We attempted to address this problem in two ways: The first approach relies on application of conditional probability measures, so as to distinguish important verb attributes for each concept. We then experimented with various probability thresholds. The second approach clusters candidate verb attributes in synonym sets, based on WordNet[11] information, so as to reduce the number of considered attributes. However, despite our efforts, hierarchical clustering outperformed FCA. Another disadvantage of FCA is the exponential time complexity $O(2^n)$, compared to the quadratic time complexity $O(N^2)$ of agglomerative clustering.

For the acquisition of non-taxonomic relations, we observe that the association rules approach in *OntoGain* is quite effective in identifying some of the most important concept relationships. The conditional probability approach attempts to establish the correct level of generalisation in the concept hierarchy. However, our results indicate that the dependency filtering process adopted in the association rules method produces highly reasonable dependencies and succeeds in pruning less significant ancestral rules and relations for modelling the domain.

4.2 Comparison to other methodologies

Systems similar to *OntoGain*, such as Text2Onto rely mostly on single-word term extraction for concept identification, thus resulting in the formation of huge lists of spurious terms and relations which are not descriptive of the domain. Our experiments with the C/NC value method for concept acquisition indicate that a method designed for the extraction of domain concepts and multi-word terms, rather than mere keywords and single-word terms, provides better conceptual representations, both in terms of detailed semantics, as well as in terms of domain coverage. Moreover, unlike information extraction type of approaches, the approach proposed in *OntoGain* for concept and relation acquisition is applied in an unsupervised and knowledge-poor manner. Another advantage of *OntoGain*

[11] http://wordnet.princeton.edu

over systems such as Text2Onto is that it outputs the results of each ontology acquisition step in OWL. The conformance of *OntoGain* output to such standards allows for easier results visualisation in any OWL compliant ontology editor, and easier ontology editing, maintenance, reuse and exchange.

In our attempt to apply the Text2Onto system to our domain corpora we were not successful. We attempted to segment our corpus but the system was running out of memory, even though we ran it on a 64-bit server reserving 3 GB of heap space. We consider that this was due to the size of input data (∼250 Mbytes for OHSUMED) which resulted in storing and processing hundreds of thousands of single-word term concepts. The subsequent effort in detection of taxonomic and non-taxonomic relations from such amounts of data leads the program to crash. This observation strengthens our assumption that multi-word terms lead to more compact representations of the examined domain, yielding dense and meaningful listings of multi-word term concepts.

5 Conclusions

We introduced *OntoGain*, a platform for ontology acquisition from texts using multi-word terms. For the acquisition of taxonomic and non-taxonomic relationships, we comparatively investigate four approaches: agglomerative hierarchical clustering and formal concept analysis, for the taxonomy development, and association rules and conditional probabilities, for the detection of non-taxonomic relations. All methods are adapted for multi-word term concept input and comparatively assessed in two different domains, the medical and the computer science domains. This evaluation indicates that agglomerative clustering and association rules outperform any other method combination reaching up to 70% precision for identification of taxonomic and non-taxonomic relations respectively in both corpora.

Investigation of method combinations for each task (e.g., agglomerative clustering with formal concept analysis, or association rules with conditional probabilities), incorporating methods for extracting or learning concept attributes (e.g., “small”, “large”, “black”, “white”), resolving term ambiguities as well as incorporating Hearst lexico-syntactic patterns for revealing additional relationships types (e.g., “part-of”) are issues for further research.

References

1. Pinto, H., Martins., J.: Ontologies: How can They be Built? Knowledge and Information Systems **6**(4) (2004) 441–464
2. Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: A Self-Organizing Framework for Information Extraction. In: In Proc. of the 18^{th} Intern. World Wide Web Conf. (WWW 2009), Madrid, Spain, ACM Press (2009) 631–640
3. Pantel, P., Pennacchiotti, M.: Automatically Harvesting and Ontologizing Semantic Relations. In: Proc. of the 2008 Conf. on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Amsterdam, The Netherlands, IOS Press (2008) 171–195

4. Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F.: Evaluation of OntoLearn, a Methodology for Automatic Learning of Ontologies. In Buitelaar, P., Cimmiano, P., Magnini, B., eds.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press, Amsterdam, The Netherlands (2005) 569–572
5. Buitelaar, P., Cimiano, P., Frank, A., Racioppa, S.: SOBA: SmartWeb Ontology-based Annotation. In: Proc. of the Demo Session at the Intern. Semantic Web Conference (ISWC), Athens GA, USA (November 2006)
6. Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research (JAIR) **24** (2005) 305–339
7. Haav, H.M.: An application of inductive concept analysis to construction of domain-specific ontologies. In: Brandenburg University of Technology at Cottbus. (2003) 63–67
8. Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. In: Proc. of the 14^{th} European Conf. on Artificial Intelligence (ECAI'00), IOS Press (August 2000) 321–325
9. Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. In Buitelaar, P., Cimiano, P., eds.: Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, Amsterdam, The Netherlands (2008) 99–104
10. Soderland, S., Mandhani, B.: Moving from Textual Relations to Ontologized Relations. In: Proc. of the 2007 AAAI Spring Symposium on Machine Reading, Menlo Park, CA, USA, AAAI Press (2007) 85–90
11. Cimiano, P., Völker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In Montoyo, A., Munoz, R., Metais, E., eds.: Proc. of the 10^{th} Intern. Conf. on Applications of Natural Language to Information Systems (NLDB). Volume 3513 of LNCS., Alicante, Spain, Springer (June 2005) 227–238
12. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press (2005)
13. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-Word Terms: The C-Value/NC-Value Method. Intern. Journal of Digital Libraries **3**(2) (2000) 117–132
14. Witschel, H.: Terminology Extraction and Automatic Indexing – Comparison and Qualitative Evaluation of Methods. In: Proc. of Terminology and Knowledge Engineering (TKE). (2005)
15. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer (2006)
16. Brank, J., Grobelnik, M., Mladenic, D.: A Survey of Ontology Evaluation Techniques. In: Proc. of the Conf. on Data Mining and Data Warehouses (SiKDD 2005), Ljubljana, Slovenia (Oct. 2005)
17. Kavalec, M., Maedche, A., Svtek, V.: Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning. In van Emde Boas, P., Pokorn, J., Bielikov, M., Stuller, J., eds.: SOFSEM. Volume 2932 of Lecture Notes in Computer Science., Springer (2004) 249–256
18. Nenadic, G., Spasic, I., Ananiadou, S.: Automatic Discovery of Term Similarities Using Pattern Mining. Intl. Journal of Terminology **10**(1) (2004) 55–80
19. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin – Heidelberg (1999)

20. Hindle, D.: Noun Classification from Predicate-Argument Structures. In: Proc. of the 28^{th} Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburgh, PA, USA (June 1990) 268–275
21. Resnik, P.: Selectional Preference and Sense Disambiguation. In: Proc. of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, DC (1997)
22. Ganter, B., Reuter, K.: Finding all Closed Sets: A General Approach. Order **8**(3) (September 1991) 283–290
23. Srikant, R., Agrawal, R.: Mining Generalized Association Rules. In: Proc. of 21^{th} Conf. on Very Large Data Bases (VLDB'95), Zurich, Switzerland, Morgan Kaufmann (September 1995) 407–419
24. Scheffer, T.: Finding Association Rules that Trade Support Optimally Against Confidence. Intelligent Data Analysis **9**(4) (2005) 381–395
25. Cimiano, P., Hartung, M., Ratsch, E.: Finding the Appropriate Generalization Level for Binary Relations Extracted from the Genia Corpus. In: Proc. of the Intern. Conf. on Language Resources and Evaluation (LREC 2006), ELRA (May 2006) 161–169
26. Hersh, W., Buckley, C., Leone, T., Hickam, D.: OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In: Proc. of the 17^{th} ACM SIGIR, Dublin, Ireland (1994) 192–201
27. Milios, E., Zhang, Y., He, B., Dong, L.: Automatic Term Extraction and Document Similarity in Special Text Corpora. In: 6^{th} Conf. of the Pacific Association for Computational Linguistics, Halifax, Canada (August 2003) 22–25