Semi-automatic construction of topic ontologies

Blaž Fortuna[1], Dunja Mladenič[1], and Marko Grobelnik[1]

Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia,
blaz.fortuna@ijs.si, dunja.mladenic@ijs.si, marko.grobelnik@ijs.si
WWW home page: http://kt.ijs.si/

Abstract. In this paper, we review two techniques for topic discovery in collections of text documents (Latent Semantic Indexing and K-Means clustering) and present how we integrated them into a system for semi-automatic topic ontology construction. The *OntoGen* system offers support to the user during the construction process by suggesting topics and analyzing them in real time. It suggests names for the topics in two alternative ways both based on extracting keywords from a set of documents inside the topic. The first set of descriptive keyword is extracted using document centroid vectors, while the second set of distinctive keyword is extracted from the SVM classification model dividing documents in the topic from the neighboring documents.

1 Introduction

When working with large corpora of documents it is hard to comprehend and process all the information contained in them. Standard text mining and information retrieval techniques usually rely on word matching and do not take into account the similarity of words and the structure of the documents within the corpus. We try to overcome that by automatically extracting the topics covered within the documents in the corpus and helping the user to organize them into a topic ontology.

A topic ontology is a set of topics connected with different types of relations. Each topic includes a set of related documents. Construction of such an ontology from a given corpus can be a very time consuming task for the user. In order to get a feeling on what the topics in the corpus are, what the relations between topics are and, at the end, to assign each document to some certain topics, the user has to go through all the documents. We try to overcome this by building a special tool which helps the user by suggesting the possible new topics and visualizing the topic ontology created so far – all in real time. This tool in combination with the corpus visualization tools [1] described in [8] aims at assisting the user in a fast semi-automatic construction of the topic ontology from a large document collection.

We chose two different approaches for discovering topics within the corpora. The first approach is a linear dimensionality reduction technique, known as Latent Semantic Indexing (LSI) [5]. This technique relies on the fact that words

[1] http://kt.ijs.si/blazf/software.html

related to the same topic co-occur together more often than words related to the different topics. The result of LSI are fuzzy clusters of words each describing one topic. The second approach we used for extracting topics is the well known k-means clustering algorithm [12]. It partitions the corpus into k clusters so that two documents within the same cluster are more closely related than two documents from two different clusters. We used these two algorithms for automatic suggestion of topics during the construction of the topic ontology.

This paper is organized as follows. Section 2 gives a short overview of the related work on building otologies. Section 3 gives an introduction to the text mining techniques we used. Details about our system are presented in Section 4, evaluation and users' feedback are presented in Section 5 followed by the future work and conclusions in Sections 6 and 7.

2 Related work on building otologies

Different approaches have been used for building ontologies, most of them using mainly manual methods. An approach to building ontologies was set up in the CYC project [6], where the main step involved manual extraction of common sense knowledge from different sources. There have been some definitions of methodology for building ontologies, again assuming manual approach. For instance, the methodology proposed in [19] involves the following stages: identifying the purpose of the ontology (purpose, intended application, range of users), building the ontology, evaluation and documentation. The building of the ontology is further divided in three steps. The first is *ontology capture*, where key concepts and relationships are identified, a precise textual definition of them is written, terms to be used to refer to the concepts and relations are identified, the involved actors agree on the definitions and terms. The second step involves *coding of the ontology* to represent the defined conceptualization in some formal languages (committing to some meta-ontology, choosing a representation language and coding). The third step involves possible *integration* with existing ontologies. An overview of the methodologies for building ontologies is provided in [7], where several methodologies, including the above described one, are presented and analyzed against the IEEE Standard for Developing Software Life Cycle Processes viewing ontologies as parts of some software product.

Recently, a number of workshops at Artificial Intelligence and Machine Learning conferences (ECAI, IJCAI, ECML/PKDD) have focused on the problem of learning ontologies. Most of the work presented there addresses one of the following: a problem of extending an existing ontology WordNet using Web documents [1], using clustering for semi-automatic construction of ontologies from parsed text corpora [2], [16], learning taxonomic, eg., "isa", [4], and non-taxonomic, eg., "hasPart" relations [15], extracting semantic relations from text based on collocations [11], extracting semantic graphs from text for learning summaries [14].

The contribution of this paper to the field is that it presents a novel approach to semi-automatic construction of a very specific type of ontology – topic ontology. Text mining techniques (e.g. clustering) were already proven successful when used at this step (e.g. [2], [16]) and in this paper we present a very tight integration of them with a manual ontology construction tool. This allows our system to offer support to the user during the whole ontology construction process.

3 Text mining techniques

Text Mining is fairly broad in its research, addressing a large range of problems and developing different approaches. Here we present only a very small subset of the available methods, namely only those that we found the most suitable for the problem addressed in this paper.

3.1 Representation of text documents

In order to use the algorithms we will describe later we must first represent text documents as vectors. We use a standard *Bag-of-Words* (BOW) approach together with *TFIDF* weighting [17]. This representation is often referred to as *vector-space model*. The similarity between two documents is defined as the cosine of the angle between their vector representations – *cosine similarity*.

3.2 Latent Semantic Indexing

Language contains many redundant information, since many words share common or similar meaning. For computer this can be difficult to handle without some additional information – background knowledge. *Latent Semantic Indexing* (LSI), [5], is a technique for extracting this background knowledge from text documents. It uses a technique from linear algebra called Singular Value Decomposition (SVD) and bag-of-words representation of text documents for detecting words with similar meanings. This can also be viewed as extraction of hidden semantic concepts or topics from the text documents.

LSI is computed as follows. First *term-document matrix* A is constructed from a given set of text documents. This is a matrix with bag-of-words vectors of documents as columns. This matrix is decomposed using SVD so that $A = USV^T$ where matrices U and V are orthogonal and S is a diagonal matrix with ordered singular values on the diagonal. Columns of matrix U form an orthogonal basis of a subspace in bag-of-words space and vectors with higher singular values carry more information. Based on this we can view vectors that form the basis as concepts or topics. The space spanned by these vectors is called *Semantic Space*.

3.3 K-Means clustering

Clustering is a technique for partitioning data so that each partition (or cluster) contains only points which are similar according to some predefined metric. In the case of text this can be seen as finding groups of similar documents, that is documents which share similar words.

K-Means [12] is an iterative algorithm which partitions the data into k clusters. It has already been successfully used on text documents [18] to cluster a large document corpus based on the document topic and incorporated in an approach for visualizing a large document collection [10]. You can see the algorithm roughly in Algorithm 1.

```
Algorithm 1: K-Means.
Input: A set of data points, a distance metric, the desired number of clusters $k$
Output: Clustering of the data points into $k$ clusters
Set $k$ cluster centers by randomly picking $k$ data points as cluster centers
repeat
    Assign each point to the nearest cluster center
    Recompute the new cluster centers
until the assignment of data points has not changed
```

3.4 Keywords extraction

We used two methods for extracting keywords from a given set of documents: (1) keyword extraction using centroid vectors and (2) keyword extraction using Support Vector Machines. We used these two methods to generate description for a given topic based on the documents inside the topic.

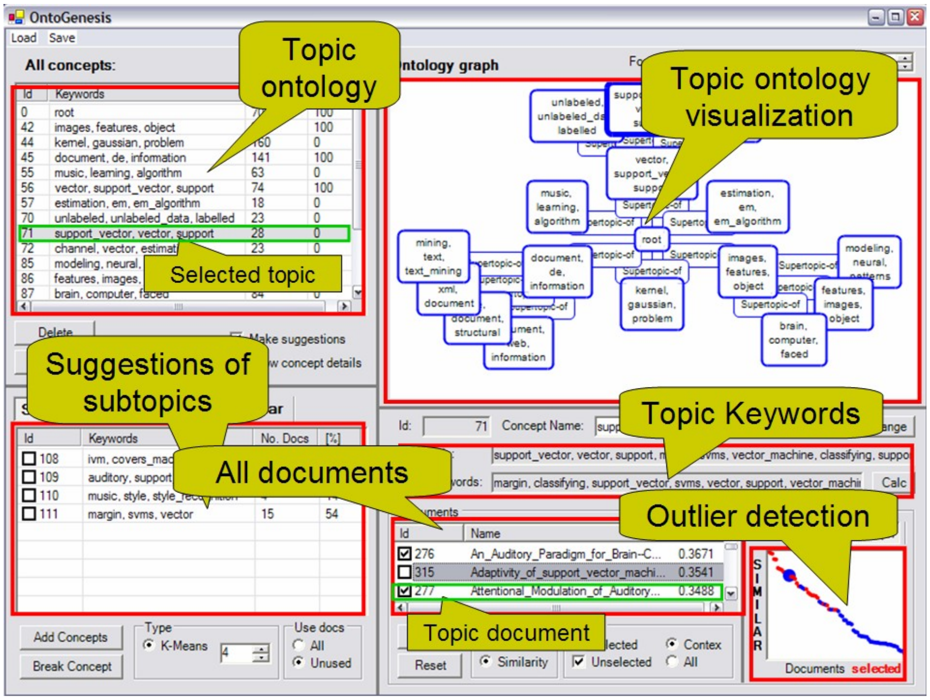
The first method works by using the centroid vector of the topic (centroid is the sum of all the vectors of the document inside the topic). The main keywords are selected to be the words with the highest weights in the centroid vector.

The second method is based on the idea presented in [3] which uses Support Vector Machine (SVM) binary classifier [13]. Let A be the topic which we want to describe with keywords. We take all the documents from the topics that have A as a subtopic and mark these documents as negative. We take all the documents from the topic A and mark them as positive. If one document is assigned both negative and positive label we say it is positive. Then we learn a linear SVM classifiers on these documents and classify the centroid of the topic A. Keywords describing the concept A are the words, which's weights in SVM normal vector contribute most when deciding if the centroid is positive (it belongs to the topic).

The difference between these two approaches is that the second approach takes into account the context of the topic. Let's say that we have a topic named 'computers'. When deciding what the keywords for some subtopic A are, the first method would only look at what the most important words within the subtopic A are and words like 'computer' would most probably be found important. However, we already know that A is a subtopic of 'computers' and we are more interested in finding the keywords that separate it from the other documents within the ‘computers’ topic. The second method does that by taking the documents from all the super-topics of A as a context and learns the most crucial words using SVM.

4 Semi-automatic construction of topic ontologies

We view semi-automatic topic ontology construction as a process where the user is taking all the decisions while the computer helps by giving suggestions for the topics, automatically assigning documents to the topics and suggesting names for the topics. The suggestions are applied only when the users decides to do so. The computer also helps by visualizing the topic ontology and the documents.



All concepts:

Id	Keywords		
0	root		
42	images, features, object		100
44	kernel, gaussian, problem	160	0
45	document, de, information	141	100
55	music, learning, algorithm	63	0
56	vector, support_vector, support	74	100
57	estimation, em, em_algorithm	18	0
70	unlabeled, unlabeled_data, labelled	23	0
71	support_vector, vector, support	28	0
72	channel, vector, estimat	23	0
85	modeling, neural,		
86	features, images,		
87	brain, computer, faced		

Id	Keywords	No. Docs	[%]
□ 108	ivm, covers_mac		
□ 109	auditory, support		
□ 110	music, style, style_rec		
□ 111	margin, svms, vector	15	54

Type: ● K-Means 4

Use docs: ○ All ● Unused

Id: 71 Concept Name: supp

support_vector, vector, support, m vms, vector_machine, classifying, suppo

words: margin, classifying, support_vector, svms, vector, support, vector_machi

Id	Name	
☑ 276	An_Auditory_Paradigm_for_Brain--C...	0.3671
□ 315	Adaptivity_of_support_vector_machi...	0.3541
☑ 277	Attentional_Modulation_of_Auditory...	0.3488

● Similarity ☑ Unselected ● Contex ○ All

Fig. 1. Screen shot of the interactive system OntoGen for construction of topic ontologies.

In Figure 1 you can see the main window of the interactive system *OntoGen* we developed. The system has three major parts that will be further discussed in following subsections. In the central part of the main window is a visualization of the current topic ontology (*Ontology visualization*). On the left side of the window is a list of all the topics from this ontology. Here the user can select the topic he wants to edit or further expand into subtopics. Further down is the list of suggested subtopics for the selected topic (*Topic suggestion*) and the list with all topics that are in relationship with the selected topic. At the bottom side of the window is the place where the user can fine-tune the selected topic (*Topic management*).

4.1 Ontology visualization

While the user is constructing/changing topic ontology, the system visualizes it in real time as a graph with topics as nodes and relations between topics as edges. See Figures 1, 2 and 3 for examples of the visualization.

4.2 Topic suggestion

When the user selects a topic, the system automatically suggests what the subtopics of the selected topic could be. This is done by LSI or k-means algorithms applied only to the documents from the selected topic. The number of suggested topics is specified by the user. Then, the user selects the subtopics he finds reasonable and the system automatically adds them to the ontology with relation 'subtopic-of' to the selected topic. The user can also decide to replace the selected topic with the suggested subtopics. In Figure 1 you can see how this feature is implemented in our system.

4.3 Topic management

The user can manually edit each of the topics he added to the topic ontology. He can change which documents are assigned to this topic (one document can belong to more topics), what is the name of the topic and what is the relationship of the topic to other topics. The main relationship "subtopic-of" is automatically induced when subtopics are added to the ontology as described in the previous section. The user can control all the relations between topics by adding, removing, directing and naming the relations.

Here the system can provide help on more levels:

- The system automatically assigns the documents to a topic when it is added to the ontology.
- The system helps by providing the keywords describing the topic using the methods described in Section 3. This can assist user when naming the topic.
- The system computes the cosine similarity between each document from the corpus and the centroid of the topic. This information can assist the user when searching for documents related to the topic. The similarity is shown on the list of documents next to the document name and the graph of similarities is plotted next to the list. This can be very practical when searching for outliers inside the concepts or for the documents that are not in the concepts but should be in considering their content.

– The system also computes similarities between the selected topic and all the other topics from the ontology. For the similarity measure between two topics it uses either the cosine similarity between their centroid vectors or the intersection between their documents.

5 OntoGen in practice

In the previous sections we described the main components of the system and the text-mining techniques behind them. Here we will show how do these components combine in a sample task of building a topic ontology and present users' feedback from cases studies which used OntoGen.

5.1 An example topic ontology

In this section we will show example of a topic ontology constructed from 7177 company descriptions taken from Yahoo! Finance[2]. Each company is described with one paragraph of text. A typical description taken from Yahoo! would look as follows:

> Yahoo! Inc. is a provider of Internet products and services to consumers and businesses through the Yahoo! Network, its worldwide network of online properties. The Company's properties and services for consumers and businesses reside in four areas: Search and Marketplace, Information and Content, Communications and Consumer Services and Affiliate Services. . .

Using OntoGen on this descriptions one can in very little time (in our case just around 15 minutes!) create an ontology of areas that companies from Yahoo! Finance cover. Companies are also automatically positioned inside this ontology. The whole ontology generated with OntoGen is show in Figure 2 and zoom into part of the topic hierarchy depicted in Figure 3. In constructions of this topic ontology all the elements of OntoGen were used and the SVM keyword extraction method was shown to be very useful when naming topics that are far from the root. We kept most of the suggestions for topics and many of them were refined with help of our visualization for the outlier detection (see the bottom right part of Figure 1). Also, relation management was found very useful since the automatically discovered relations are not always optimal. By spending more time this ontology could be further developed to cover the ares in more details.

5.2 Case studies and users' feedback

The presented system OntoGen was used for semi-automatic ontology construction in several case studies for modeling topic ontologies of the following domains:

[2] http://finance.yahoo.com/

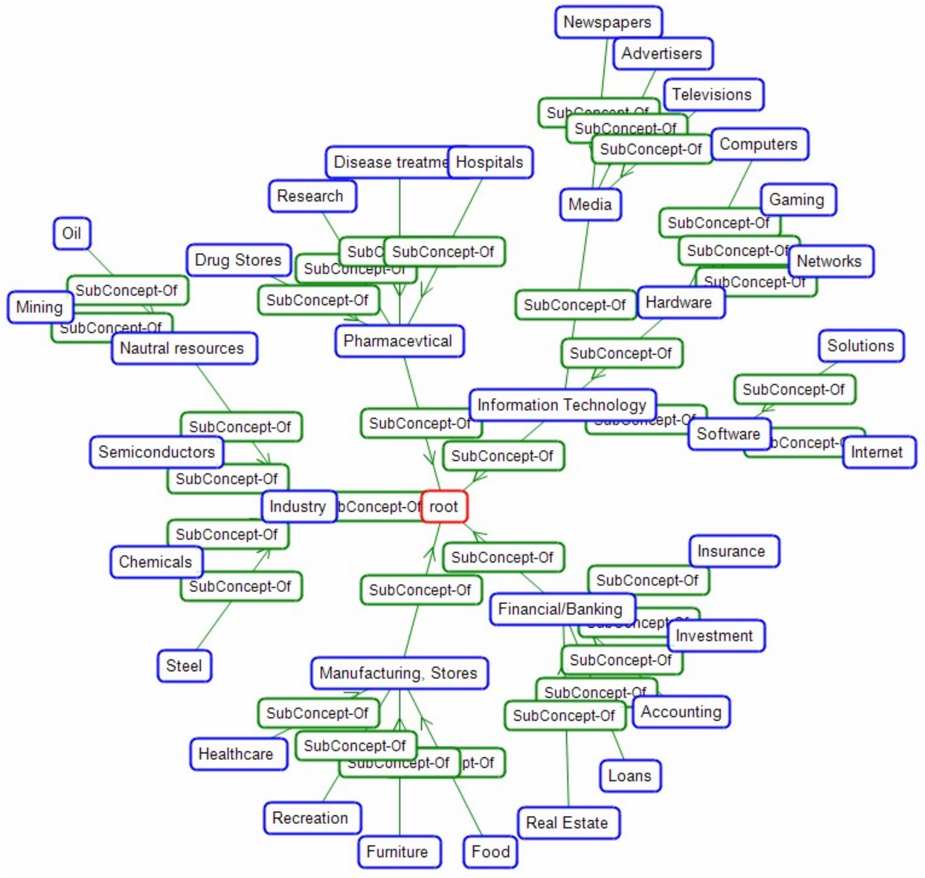


Fig. 2. Topic Ontology constructed from company descriptions. The top node of topic ontology is located in the center of the figure.

- legal judgements in “Spanish legal case study” inside European project SEKT,
- virtual organizations inside inside European project ECOLEAD and
- news articles published by Slovene Press Agency (STA).

In all the three cases, the users were in fact domain experts, knowledgable about the domain but had little experience in knowledge engineering and practically no experience in machine learning. They were fast at learning how to use the system and were in general very pleased with its performance and the amount of time they needed to derive a desirable results. The domain experts were also satisfied with the final topic ontologies constructed for the all three cases.

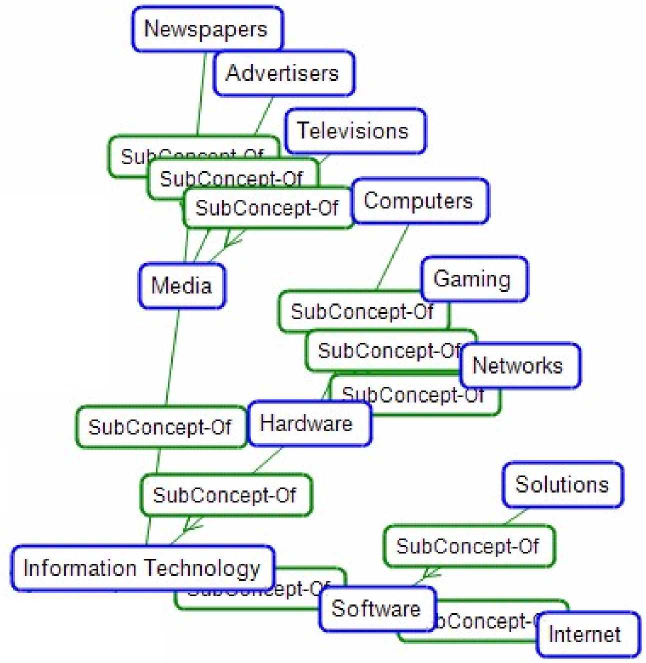


Fig. 3. Zoom in to *Information Technology*part of Yahoo! Finance topic ontology.

Many of the comments we got as a feedback from the users were related to the user interface of the system. For illustration, we list here the most interesting comments:

- no *undo* function,
- too little information is presented about suggested topics,
- editing of document membership for a specific topic is unclear and
- more interactive topic ontology visualization (folding, zooming).

Some other comments were more closely related to the topic suggestion and keyword extraction methods:

- "I would like to mark a keyword *not relevant* so the system will ignore it when generating suggestions?"
- "I know the name of the topic I would like to add to the topic ontology but the system does not find it."

These comments show the limits of the suggestion methods currently included in the system and they were of great help to use when deciding what other text-mining methods to include in the future versions of the system.

We found the comments from the users to be very informative and constructive and most of them will be implemented in the next versions of OntoGen.

6 Future work

Currently we are working in two directions of adding functionality to OntoGen. The version presented in this paper can only help at discovering the topics but has no support for identification and naming of relations. The idea here is to use machine learning and text mining to discover possible relations between topics. The second direction is to include methods for incorporating background knowledge into the topic discovery algorithms [9]. This would enable building of different ontologies based on the same data. For example, the same document-database in a company may be viewed differently by marketing, management, and technical staff.

Another possible direction would be making the whole process more automatic and reduce the need for user interaction. This involves things like calculating the quality of topics suggested by the system, more automated discovery of the optimal number of topics, improved support for annotating the documents with the topics, discovering different kinds of relations between topics etc.

7 Conclusions and discussion

In this paper we presented our approach to the semi-automatic construction of topic ontologies. In the first part of the paper we presented text mining techniques we used: two methods for discovering topics within the corpus, LSI and K-Means clustering, and two methods for extracting keywords. In the second part we showed how we integrated all these methods into an interactive system for constructing topic ontologies. The system was successfully tested and used in three case studies with very satisfactory results both in terms of final results and the feedback we got from the end-users.

Even though the system was primarily designed for constructing topic ontologies it can be generalized for other types of ontologies where the instances can be described by some relevant features. In case of topic ontologies the instances are documents which are described by words as features, but it might as well be users described by products they bought or movies they saw, images described by SIFT features, etc. Clustering can still be used as a method for discovering concepts but naming the concepts can be little more trickier for cases when features are harder to understand and are not words (for example, SIFT features used). In that cases methods for keyword extraction presented in this paper would not be sufficient.

8 Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under 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., Martinez. D. *Enriching Very Large Ontologies Using the WWW.* In Proceedings of the Ontology Learning Workshop, The 14th European Conference on Artificial Inteligence (ECAI), Berlin, Germany, 2000.
2. Bisson, G., Nedellec, C., Canamero L. *Designing clustering methods for ontology building: The MoK workbench.* In Proceedings of the Ontology Learning Workshop, The 14th European Conference on Artificial Inteligence (ECAI), Berlin, Germany, 2000.
3. Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D. *Feature selection using support vector machines.* In Proceedings of the 3rd International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields, Bologna, Italy, 2002.
4. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: *Learning Taxonomic Relations from Heterogeneous Evidence.* In Proceedings of the Ontology Learning and Population Workshop, The 16th European Conference on Artificial Inteligence (ECAI), Valenci, Spain, 2004.
5. Deerwester, S., Dumais, S., Furnas, G., Landuer, T., Harshman, R.: *Indexing by Latent Semantic Analysis.* Journal of the American Society of Information Science, vol. 41, no. 6, 391-407, 1990.
6. Douglas B. Lenat R. V. Guha: *Building Large Knowledge-Based Systems* Addison Wesley, Reading, Massachusetts, 1990.
7. Lpez, M.F. *Overview of the methodologies for building ontologies.* In Proceedings of the Ontologies and Problem-Solving Methods Workshop, The 16th International Joint Conference on Artificial Inteligence (IJCAI), Stockholm, Sweden, 1999.
8. Fortuna, B., Grobelnik, M., Mladenic, D. *Visualization of text document corpus.* Informatica, vol. 29, 497-502, 2005.
9. Fortuna, B., Grobelnik, M., Mladenic, D. *Background Knowledge for Ontology Construction.* Poster at 16th International World Wide Web Conference (WWW2006), Edinburgh, Scotland, 2006.
10. Grobelnik, M., And Mladenic, D. *Efficient visualization of large text corpora.* In Proceedings of the 17th TELRI seminar, Dubrovnik, Croatia, 2002.
11. Heyer, G., Luter, M., Quasthoff, U., Wittig, T., Wolff, C. *Learning Relations using Collocations.* In Proceedings of Workshop on Ontology Learning, The 17th International Joint Conference on Artificial Inteligence (IJCAI), Seattle, USA, 2001.
12. Jain, A. K., Murty, M. N., Flynn, P. J. *Data Clustering: A Review.* ACM Computing Surveys, vol 31. no. 3, 264-323, 1999.
13. Joachims, T. *Making large-scale svm learning practical.* In Scholkopf, B., Burges, C., Smola, A., Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA, 1998.
14. Leskovec, J., Grobelnik, M., Milic-Frayling, N. *Learning Semantic Graph Mapping for Document Summarization.* In Proceedings of Workshop on Knowledge Discovery and Ontologies, 15th European Conference on Machine Learning (ECML) and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Pisa, Italy, 2004.
15. Maedche, A., Staab, S. *Discovering conceptual relations from text.* In The 14th European Conference on Artificial Inteligence (ECAI), 321-325, Berlin, Germany, 2000.

16. Reinberger, M-L., Spyns, P. *Discovering Knowledge in Texts for the learning of DOGMA-inspired ontologies.* In Proceedings of the Ontology Learning and Population Workshop, The 16th European Conference on Artificial Inteligence (ECAI), Valenci, Spain, 2004.
17. Salton, G. *Developments in Automatic Text Retrieval.* Science, vol. 253, pages 974-979, 1991.
18. Steinbach, M., Karypis, G., Kumar, V.): *A comparison of document clustering techniques.* In Proceedings of KDD Workshop on Text Mining, 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Boston, USA, 2000.
19. Uschold, M. *Towards a Methodology for Building Ontologies.* Workshop on Basic Ontological Issues in Knowledge Sharing, The 14th International Joint Conference on Artificial Inteligence (IJCAI), Motnreal, Canada, 1995.