Simple Classification into Large Topic Ontology of Web Documents

Marko Grobelnik and Dunja Mladenić

Jozef Stefan Institute, Ljubljana, Slovenia

The paper presents an approach to classifying Web documents into large topic ontology. The main emphasis is on having a simple approach appropriate for handling a large ontology and providing it with enriched data by including additional information on the Web page context obtained from the link structure of the Web. The context is generated from the in-coming and out-going links of the Web document we want to classify (the target document), meaning that for representing a document we use, not only text of the document itself, but also the text from the documents pointing to the target document, as well as the text from the documents the target document is pointing to. The idea is that providing enriched data is compensating for the simplicity of the approach while keeping it efficient and capable of handling large topic ontology.

Keywords: classification of documents, topic ontology of Web documents, Web document context, link structure of the Web.

1. Introduction

Most of the existing ontologies were developed with considerable human efforts. Examples are Yahoo! and DMoz topic ontologies containing Web pages or MESH ontology of medical terms connected to Medline collection of medical papers. We propose a simple approach to classifying documents into an existing topic ontology based on the k-nearest neighbor algorithm [4]. Documents are represented as feature-vectors, with features being single words and sequences of up to 3 consecutive words, removing rare features.

The problem of classifying documents into hierarchy of topic-classes (topic ontology) was already addressed by several researchers. An approach to classifying text documents into very simple topic ontology was presented in [2]. They used the Reuters news collection and a manually constructed topic ontology of them. In that ontology all the documents were placed at the bottom of the ontology, in the leaves representing the most specific topics. Documents are represented as Boolean word- vectors with features representing words selected using greedy algorithm that eliminates features one by one using Cross entropy measure. They compared several learning algorithms and learn document category from the hierarchical structure, dividing classification task into a set of smaller problems corresponding to the splits in the classification hierarchy nodes. They give results on three domains each having a small, 3- level topic ontology that is based on up to 1,000 documents having in total 12 nodes.

Some researchers work on larger datasets involve existing topic ontology of Yahoo! Web pages and US patent database. An approach was developed on extracting expertise from the two bottom layers of the ontology [3]. Another related approach [1] has been developed with the goal of organizing large text databases into hierarchical topic taxonomies. Similar as in [5], for each internal node in the topic taxonomy, a separate subset of features is calculated and used to build a classifier for that node. In that way, the feature set used in the classification of each node is relatively small and changes with its context. Opposite to the work reported in [5], and similar to the work of [2], the assumption is that all the documents are placed in the leaves of an ontology and a multilevel classifier is used to classify a new document to a leaf node. Namely, a new document is classified from the top of the taxonomy and, based on the classification outcome in the inner nodes, it proceeds down the taxonomy. However, these approaches do not address the problem of documents being instances of an arbitrary node rather than just ontology leaves.

Another approach developed on the Yahoo! topic ontology [5, 6] considers situations when instances are placed in any node of the topic ontology (not just its leaf nodes). Namely, it turned out that some documents were manually placed in the non-leaf nodes as their content was too general for any of the existing leaf nodes (and probably not specific/frequent enough to trigger introduction of a new concept). For a new document, the learned model returns the probability that the document is an instance of a topic, for each topic from the ontology (and the corresponding set of keywords describing the path from the root node to the topic node). The ontology structure is handled so that for each topic a separate sub-problem is defined. A set of positive and negative examples for each sub-problem are constructed from the sub-tree of the topic node. Learning is performed using the naive Bayesian classifier on the selected feature subset and the final result of learning is a set of specialized classifiers, each based only on a small subset of features. Our approach is similar in defining a sub-problem for each ontology node, but we use a simple algorithm for classification and handle much larger data sets.

2. Approach Description

To construct a classifier into topic ontology we assume that the ontology is hierarchical and contains only is-a relations with more general topics being higher in the hierarchy. There are typically two approaches to building a classifier into hierarchy: (1) flattening of the structure and building separate classifier for each class in the hierarchy/ontology — the final classification is produced from some kind of voting schema, and (2) hierarchical classifier which is appropriate just for taxonomic ontologies where for each node there is a separate classifier deciding which branch from the node should be followed in order to classify a new instance. The solution (1) is more general and allows addressing also non-taxonomic structures, but is computationally more expensive (because in the classification phase it addresses all the classifiers). Solution (2) is more efficient, because it addresses just the number of classifiers which is the logarithm of the number of classes in the ontology. Solution (1) is interesting also because it addresses each individual concept in the ontology separately, which is not the case with the solution (2), where in the cases of large taxonomic ontologies (such as DMoz), the information about the lower level concepts is lost for the higher level classifiers (which have only very broad view to the distributions about the data in lower branches).

We decided to use the solution (1) which proved to be very efficient in the combination with k-Nearest Neighbor algorithm, because it was able to classify several new documents per second into 400,000 classes (DMoz hierarchy).

As already mentioned, we have based our work on our previous work on Yahoo! topic ontology of Web pages [6]. The same as there, instances in the ontology are *html* documents, cleaned and represented as feature-vectors. We have preprocessed documents by removing stop-words using the standard set of English stop-words (525), and stemming the words using the Porter stemmer. We have also generated new features containing frequent phrases defined as n consecutive words, with n=3 and minimal frequency being 5. The whole problem is divided into sub-problems, one for each concept (topic) of the topic ontology. In order to be able to handle large topic ontologies, such as DMoz (having several million of documents and several hundred thousands concepts), we have used a simple approach to modeling based on k- nearest neighbor algorithm which gives good results in terms of accurate classifications and computational efficiency. The final model is provided in the form of a set of specialized classifiers. These classifiers are used when a new document needs to be classified into the topic ontology — we use simple k-nearest voting scheme.

Since we have used the flattened version of DMoz topic ontology, the approach is similar to classifying into flat set of classes with k-nearest neighbor. Currently we are using the most simple way of flattening — namely, taking all the documents from the sub-tree and using them to create a standard centroid vector. In the future we plan to experiment with more sophisticated strategies to create the representative vector of the class — this would include different weighting schemes for sub-trees, using information also from non-taxonomic links.

An important innovation and contribution of this paper is extending the target document with its context from the web (snippets of the pages pointing to the target document page and, a set of snippets of related documents). In general, it is possible to use other "biased" functions, the problem is only potentially high number of parameters which would need to be trained for such a "biased model". This would be possible for smaller ontologies (several tens of nodes), but is less practical for large ontologies such as DMoz (400,000 classes).

3. Architecture

The architecture of the approach is shown in Figure 1. It consists of the following steps.

1. First, we download the DMoz data from the address `http://rdf.dmoz.org/`. The data is available as two large RDF/XML files, giving the ontology structure (skeleton of the hierarchy) and the ontology contents (documents manually indexed into the ontology classes).
2. Since manipulation with large RDF files (approx 2Gb) does not allow for efficient manipulation with the data, we transform the downloaded data with a special utility DMoz2Bin into a binary form — the whole (or part of) DMoz ontology is saved on the file of approx 1Gb. The file represents binary serialized C++ object which has an internal organization allowing querying and traversing the ontology structure and data in a very efficient way. For illustration, let us just say that the ontology structure is represented as a labeled graph, all the strings are represented in a string pool, all the vectors are saved in a vector pool etc., which enables efficient storage while preserving manipulatability of the data.
3. In the next step, the efficient binary structure from the previous step is used to represent the documents using a well known bag-of-words document representation, were a feature-vector is created for each document. Each feature is represented by its TFIDF weight, as commonly used in document classification. This is performed with the special utility DMoz2Bow, which creates two files from the ontology documents: ".Bow" file with bag-of-words vectors and ".Bow-Part" with the mapping (classification) of the vectors into the structure. In this representation, we calculate for each node in the topic ontology a centroids vector of all the documents (short documents describing indexed web pages) in the node itself and its sub-tree. In other words — each bag-of-words unit represents a union of all the documents belonging to the concept and its sub-concepts.
4. The data prepared in the pervious steps is used for classification. The classification

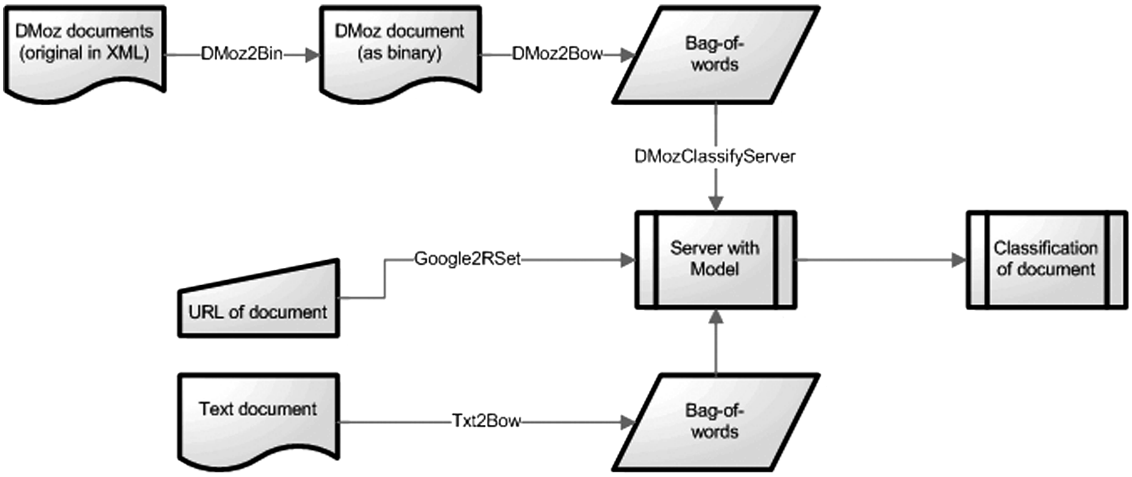


Fig. 1. Architecture of the system for extracting human expertise from DMoz topic ontology.

model consists of a set of vectors, each representing a single node from the topic ontology. Classification of a new document into the topic ontology consists of representing the document as feature-vector by transforming text of the document into the bag-of-words representation, as already described using a special utility Txt2Bow, then finding the concepts whose centroids vector is the most similar to the target document. For calculating similarity between the document and the concept centroids vector we use the standard cosine measures on TFIDF feature-vectors.

5. If the target document which we want to classify has also a URL address, we enrich the document with its context consisting of two parts: snippets of the pages pointing to our target document (using link Google function) and snippets of the pages which are related to our target page (Google target page) using a special utility Google2RSet.
6. Once the model in constructed, it can be used for classification of new documents. We have developed DMoz classification server (DMozClassifyServer) which loads the model into the memory enabling for efficient k-nearest neighbor classification. The software offers functionality as web user interface or as a web service (providing results in XML format). On the input, the user provides a URL (if available) of the target document and the text of the document. On the output, we get a list of the most probable categories (concepts) from DMoz topic ontology with associated weights and a list of the most probable keywords (calculated from the path segments from the names of the DMoz categories).

4. Data Characteristics

In this paper we are proposing an approach to efficient classification into large topic ontology tested on DMoz. The ontology contains 15 levels with the following number of topics (nodes) at each level, provided in the form (level number : number of topics at that level): 1 : 1, 2 : 17, 3 : 556, 4 : 6778, 5 : 30666, 6 : 60434, 7 : 78713, 8 : 88864, 9 : 70981, 10 : 39436, 11 : 22736, 12 : 5258, 13 : 1318, 14 : 126, 15 : 5. For instance, at the first level, there is only a root node with one topic — 'DMoz'. At level two there are 17 topics: '*Arts*', '*Business*', '*Computers*', '*Games*', '*Health*' etc. At level three there are 556 topics, at level four 6778 topics,... If we look into the inner structure of the ontology (non-leaf nodes), there is 94113 nodes having at least one branch. The average number of branches of an ontology node is 4.31277 (with the standard deviation of 8.26148), the first quartile is 1, the mediana is 2, the third quartile is 4, the deciles are as follows: Dec0 : 1 Dec1 : 1 Dec2 : 1 Dec3 : 1 Dec4 : 1 Dec5 : 2 Dec6 : 3 Dec7 : 4 Dec8 : 5 Dec9 : 10 Dec10 : 398. From that statistics we can see that over 40% of the ontology nodes have only one branch, 70% have up to four branches, while only less than 10% have over 10 branches. The minimum number of branches in the whole ontology is 1 and the maximum is 398.

If we look into the number of external URLs that the DMoz categories are pointing to, there are 405889 different categories, some of them have no pointers to external URLs, the maximum number of poitners to the outside Web pages is 12301, the mean number is 8.32918 (with the standard deviation 28.9138) URLs per category. The first quartile is 1, the mediana is 3, the third quartile is 8, the deciles are as follows: Dec0 : 0 Dec1 : 0 Dec2 : 1 Dec3 : 1 Dec4 : 2 Dec5 : 3 Dec6 : 5 Dec7 : 7 Dec8 : 11 Dec9 : 19 Dec10 : 12301. We can see that over 30% of categories have only one link to the outside pages, over 60% to five pages and less than 10% of categories point to over 19 pages.

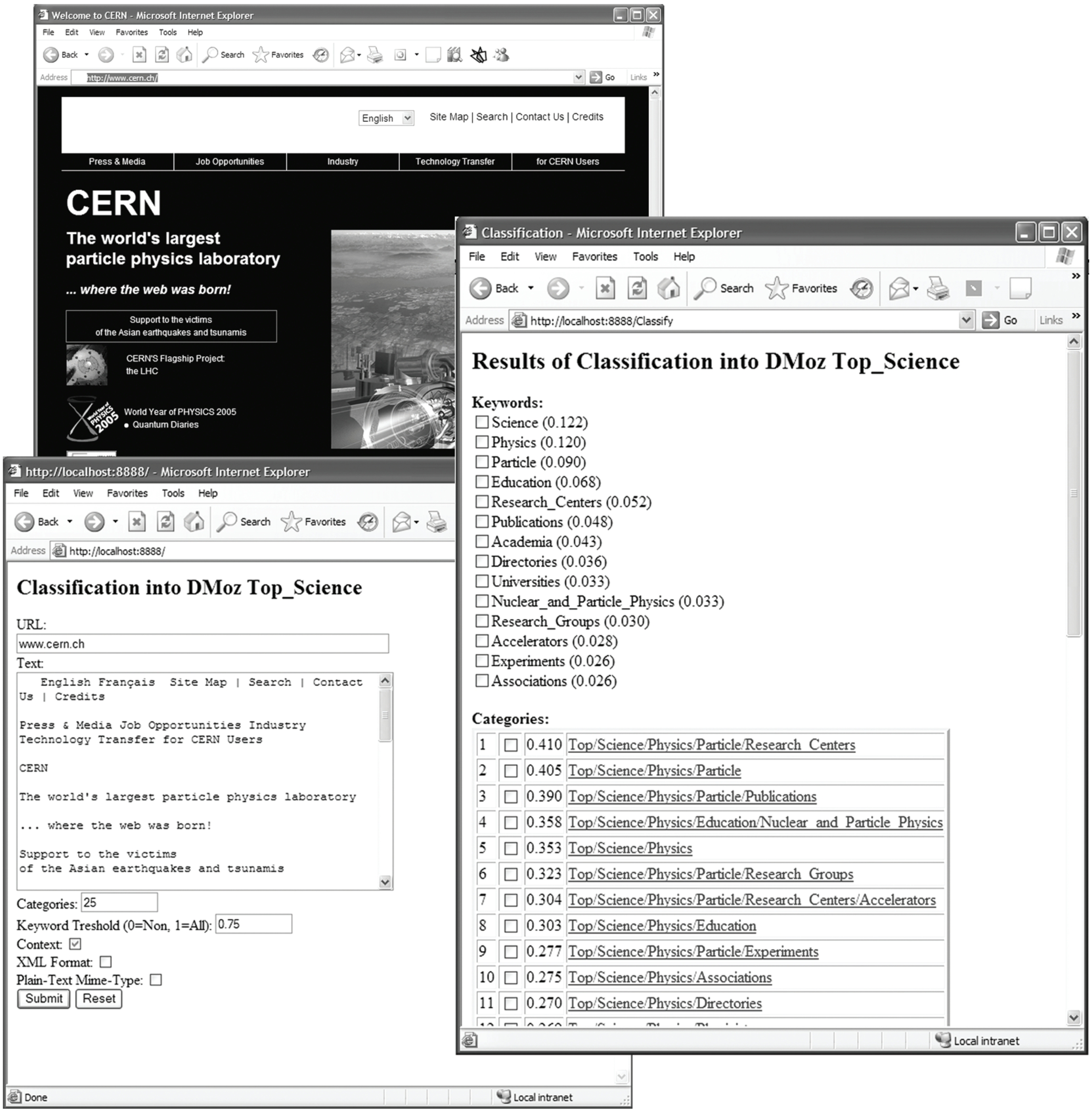
5. Conclusions and Future Work

The paper proposes a simple approach for classifying Web documents into large topic ontology. In addition to using the content of the document to be classified, we are also using information on the Web page context obtained from the link structure of the Web. For illustration of the system, let us take a look at an example usage of the system, where '*Science*' part of DMoz is used and we are classifying CERN institute home-page.

First we generate a model using the documents already classified into DMoz topic ontology under '*Science*' (performing steps 1, 2 and 3, as described in Section 3). Then, we run the server on the '*Science*' part of DMoz using the generated model (step 6 from Section 3) and provide it with the Web page that we want to classify (CERN institute home page). We access the classification server via the Web browser, type in the URL of the page and copy the text from the target page (see Figure 2). Notice that we can use the classification also in the case when one of the information (either URL or text) is missing.

After pressing the "Submit" button, the document is classified (performing steps 4 and 5 from Section 3) and the server sends back the list of keywords which are the most relevant for the submitted page (in our illustrative example in Figure 2: "Science, Physics, Particle, Education, Research_Centers,. . .,) and the list of the most relevant DMoz categories — ontology nodes (Top/Science/Physics/Particles/Research_Centers, Top/Science/Physics/Particles/,. . ., see Figure 2).

In the future development we plan to first extensively evaluate the approach for classification of new documents into the topic ontology of different size, evaluate the benefit of extend-


Welcome to CERN - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Links

http://www.cern.ch/

English

Site Map | Search | Contact Us | Credits

Press & Media | Job Opportunities | Industry | Technology Transfer | for CERN Users

CERN

The world's largest particle physics laboratory

... where the web was born!

Support to the victims of the Asian earthquakes and tsunamis

CERN'S Flagship Project: the LHC

World Year of PHYSICS 2005

- Quantum Diaries

http://localhost:8888/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address

http://localhost:8888/

Classification into DMoz Top_Science

URL: www.cern.ch

Text:

```
   English Français  Site Map | Search | Contact
Us | Credits

Press & Media Job Opportunities Industry
Technology Transfer for CERN Users

CERN

The world's largest particle physics laboratory

... where the web was born!

Support to the victims
of the Asian earthquakes and tsunamis
```

Categories: 25

Keyword Treshold (0=Non, 1=All): 0.75

Context: ☑

XML Format: □

Plain-Text Mime-Type: □

Submit
Reset
Done
Local intranet
Classification - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go
Links

http://localhost:8888/Classify

Results of Classification into DMoz Top_Science

Keywords:

□ Science (0.122)
□ Physics (0.120)
□ Particle (0.090)
□ Education (0.068)
□ Research_Centers (0.052)
□ Publications (0.048)
□ Academia (0.043)
□ Directories (0.036)
□ Universities (0.033)
□ Nuclear_and_Particle_Physics (0.033)
□ Research_Groups (0.030)
□ Accelerators (0.028)
□ Experiments (0.026)
□ Associations (0.026)

Categories:

1	□	0.410	Top/Science/Physics/Particle/Research_Centers
2	□	0.405	Top/Science/Physics/Particle
3	□	0.390	Top/Science/Physics/Particle/Publications
4	□	0.358	Top/Science/Physics/Education/Nuclear_and_Particle_Physics
5	□	0.353	Top/Science/Physics
6	□	0.323	Top/Science/Physics/Particle/Research_Groups
7	□	0.304	Top/Science/Physics/Particle/Research_Centers/Accelerators
8	□	0.303	Top/Science/Physics/Education
9	□	0.277	Top/Science/Physics/Particle/Experiments
10	□	0.275	Top/Science/Physics/Associations
11	□	0.270	Top/Science/Physics/Directories

Fig. 2. The example Web page, the system accessed via Web server and the classification results into Science part of DMoz.

ing the document representation by the context from the Web and compare our approach to some of the existing approaches to hierarchical classification. In addition to experimenting on DMoz topic ontology, we would like to test our approach on some other datasets, where the context can be obtained from different sources, not necessarily containing Web documents.

6. Acknowledgement

This work was supported by the Slovenian Ministry of Education, Science and Sport and the IST Programme of the European Community under SEKT Semantically Enabled Knowledge Technologies (IST-1-506826-IP), ALVIS Superpeer Semantic Search Engine (IST-1-002068-STP), and PASCAL Network of Excellence (IST-2002- 506778). This publication only reflects the authors' views.

References

[1] S. CHAKRABARTI, B. DOM, R. AGRAWAL, P. RAGHAVAN, (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal* (1998) pp. 7: 163–178, Spinger-Verlag 1998.

[2] D. KOLLER, M. SAHAMI, (1997). Hierarchically classifying documents using very few words, *Proceedings of the 14th International Conference on Machine Learning* ICML-97, pp. 170–178, Morgan Kaufmann, San Francisco, CA.

[3] A. MCCALLUM, R. ROSENFELD, T. MITCHELL, A. NG, (1998). Improving Text Classification by Shrinkage in a Hierarchy of Classes, *Proceedings of the 15th International Conference on Machine Learning* ICML-98, Morgan Kaufmann, San Francisco, CA.

[4] T. M. MITCHELL, (1997). Machine Learning. The McGraw-Hill Companies, Inc.

[5] D. MLADENIĆ, M. GROBELNIK, (2003). Feature selection on hierarchy of web documents. *Journal of Decision Support Systems*, 35, pp. 45–87.

[6] D. MLADENIĆ, M. GROBELNIK, (2004). Mapping documents onto web page ontology. In: Web mining: from web to semantic web (B. Berendt, A. Hotho, D. Mladenić, M. W, Van Someren, M. Spiliopoulou, G. Stumme, eds.), Lecture notes in artificial intelligence, Lecture notes in computer science, vol. 3209, Berlin; Heidelberg; New York: Springer, 2004, pp. 77–96.

Recived: June, 2005.
Accepted: October, 2005.

Contact address:
Marko Grobelnik
Jozef Stefan Institute
Jamova 39
Ljubljana
Slovenia
e-mail: Marko.Grobelnik@ijs.si

Dunja Mladenić
Jozef Stefan Institute
Jamova 39
Ljubljana
Slovenia
e-mail: Dunja.Mladenic@ijs.si
http://kt.ijs.si/Dunja/

MARKO GROBELNIK is an expert in analysis of large amounts of complex data with the purpose to extract useful knowledge. In particular, the areas of his expertise comprise: data mining, text mining, information extraction, link analysis, and data visualization as well as more integrative areas such as semantic Web, knowledge management and artificial intelligence. Apart from research on theoretical aspects of unconventional data analysis techniques, he has valuable experience in the field of practical applications and development of business solutions based on the innovative technologies. Marko was employed as a researcher first, at the Computer Science Department at the University of Ljubljana and later at the Department of Knowledge Technologies at Jozef Stefan Institute, Ljubljana, the main national research institute for natural sciences in Slovenia. His primary focus of research and applications is intelligent data analysis which deals with unconventional scenarios going beyond classical statistical approaches and solving problems including recently unstructured or semi structured data. His main achievements are from the field of Text-Mining (analysis of large amounts of textual data), having a leading role in scientific and applicative projects funded by European Commission, having projects with industries including Slovenian publishing companies and Slovenian National and University Library and enterprises such as Microsoft Research. He has published several papers in refereed conferences and journals and served in program committees of several international conferences and organized a series of international events in the area of text mining, link analysis and data mining.

Marko Grobelnik is on the Management Board of several EU project including 6FP IP project "SEKT — Semantically-Enabled Knowledge Technologies" (2004–2006), 6FP NoE project "PASCAL - Pattern Analysis, Statistical Modelling and Computational Learning" (2004–2007), 6FP STREP project "ALVIS — Superpeer Semantic Search Engine" (2004–2006). He has two bilateral projects with Microsoft Research, Cambridge, UK on "Application of Advanced Natural Language Processing to Text Mining and Summarization" (2002–2003) and on "Text Analysis using Natural Language Processing" (2000–2001). He is also intensively involved in several national projects including "Construction of Archive for Slovenian Web Publications", a joint project with National and University Library of Slovenia (2002–2004), "Design and Analysis of Slovenian Digitalized Electronic Publications of National Importance", a joint project with National and University Library of Slovenia (2002–2004), "Language Resources for the Slovene Language", a joint project with University of Ljubljana (Faculty of Arts, Faculty of Social Studies), DZS d.d. publishing company and a software development company Amebis d.o.o. (2002–2005).

DUNJA MLADENIC is an expert on study and development of machine learning, data mining and text mining techniques and their application on real-world problems from different areas such as publishing, medicine, pharmacology, manufacturing, economy. Her current research focuses on data analysis, with particular interest in learning from Text and the Web including personal intelligent agents. She works as a researcher at the Department of Knowledge Technologies of the J. Stefan Institute, Ljubljana, Slovenia since 1992. She graduated in computer science from the University of Ljubljana and continued as a PhD student focused on Artificial Intelligence. She got her MSc and PhD in computer science at the University of Ljubljana in 1995 and 1998 respectively. She worked at School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, as a visiting researcher in 1996–1997 and in 2000–2001.

Dunja Mladenic was coordinating EU 5^{h} RTD FP project "Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (Sol-Eu-Net)" (2000–2003). She is on the Management Board of several EU project including 5FP NoE projects "KD-Net — The European Knowledge Discovery Network of Excellence" (2002–2004) and KMForum (2001–2003), 6FP IP project "SEKT — Semantically — Enabled Knowledge Technologies" (2004-2006), 6FP NoE project "PASCAL — Pattern Analysis, Statistical Modelling and Computational Learning" (2004–2007), 6FP SSA project "CEC-WYS: Central European Centre for Women and Youth in Science" (2004–2006). She is the Slovenian member of the EC Enwise Expert Group "Promoting women scientists from the Central and Eastern European countries and the Baltic States to produce gender equality in science in the wider Europe". She serves as project evaluator of project proposals for EC programme on Information and Society Technology (IST). In 2001, she was evaluator of project proposals for National Science Foundation (NSF) initiative on Information Technology Research (ITR), NSF 00–126, USA.

She has published several papers in refereed conferences and journals, served in the program committees of different international conferences and organized several international events in the area of text mining, link analysis and data mining. She is co-editor of the book "Data Mining and Decision Support: Integration and Collaboration" (Mladenic, Lavrac, Bohanec, Moyle, eds.) Kluwer Academic Publishers, 2003 and of the book "Web Mining: from Web to Semantic Web" (Berendt, Hotho, Mladenic, Someren, Spiliopoulou, Stumme, eds.) Lecture notes in AI, Berlin; Heidelberg; New York: Springer, 2004.