28. Mednarodna konferenca o razvoju organizacijskih znanosti
NOVE TEHNOLOGIJE, NOVI IZZIVI
25. - 27. marec 2009, Portorož, Slovenija

Uporaba orodij za rudarjenje v besedilih pri odkrivanju novih povezav v literaturi

Text mining for uncovering new relations in literature

Bojan Cestnik
Temida d.o.o., Ljubljana, Slovenija
Jozef Stefan Institute, Ljubljana, Slovenia
bojan.cestnik@temida.si

Povzetek
V članku najprej na kratko predstavimo tehnologijo rudarjenja v besedilih. Nato podajamo in opišemo glavna področja uporabe, pri čemer posebej izpostavimo vlogo učenja ontologij iz besedil. Predstavitev naučenih konceptov v obliki ontologije je osnova za strukturiranje znanja na izbranem problemskem področju in omogoča boljše sporazumevanje s strokovnjaki s posameznega področja. Ontologije so primerne tudi za učinkovito seznanjanje s problemskim področjem, saj vsebujejo posplošen pogled na obravnavano množico besedil. Uporabnost tehnologije rudarjenja v besedilih podrobneje ilustriramo na dveh primerih: pri odkrivanju novih povezav v literaturi in pri izboljšanju razumljivosti besedil.

Ključne besede: rudarjenje v besedilih, predstavitev znanja, ontologije, vizualizacija

Abstract
In the paper we first give an overview of the literature mining technology. Then we list its main application areas with a special emphasis on the role of learning ontologies from texts. Representation of learned concepts in the form of ontology is a basis for structuring domain knowledge and facilitates better communication with domain experts. Ontologies are suitable also for effective apprehension of a target problem domain, since they comprise generalized view on the studied literature. To further illustrate the utility of literature mining we focus on two specific case studies: discovering new connections between domain concepts and improving the comprehension of legal documents.

Keywords: literature mining, knowledge representation, ontologies, visualization

1 Introduction

The quantity of knowledge stored in the form of scientific publications, articles and books reveals an exponential growth in the recent years. Consequently, the interest in using text-mining technologies evidently rises, primarily because effectively following the progress in large quantity of published knowledge is almost impossible even in a relatively narrow field of interest (e.g. medicine). Written text typically contains large quantity of information that is usually encoded in a way hard to comprehend for automatic approaches. In spite of that fact, information technology spawned, besides the area of text mining, several important application areas like information retrieval, computational linguistics, text categorization, ontology learning and hypothesis formation. In the following paragraphs each of the enumerated areas are shortly described and compared to text mining.

The information retrieval pursues a goal of helping to find documents that correspond to user's search requests and criteria (Singhal, 2001). Its main application area is document searching over the Internet. The main difference between text mining and information retrieval lies in the fact that the text mining endorses discovery of new knowledge and searches for typical patterns within collections of text documents, while information retrieval only effectively searches text documents to identify those that meet user's search criteria. The later, therefore, does not imply searching for new knowledge; it merely focuses on displaying the information that is already present in a collection of documents, but is difficult to find because of the sheer size of the collection.

Similarly to text mining, the field of computational linguistics includes searching for useful patterns in large collections of text documents (usually called text corpora) based on statistical techniques. The difference between the two fields lies in the nature of target patterns. A typical pattern that is of interest for the computational linguistics field contains a list of words that appear in a given document subset more frequently. However, the user expectations from the text-mining field usually exceed the options that are available within the computational linguistics (Hutchins, 1999).

The field of text categorization (Sebastiani, 2002) also shares several similarities with text mining. Categorization implies classification of input documents into one of the predefined categories or classes. Since these classes are statically determined in advance, there is no new knowledge involved in the process. On the other hand, several modern approaches for text categorization contain also methods for detecting unexpected patterns in text, which is similar to text mining.

Ontologies are used for structured presentation of concepts and their relations in several areas of our lives. In science, they have been used for decades to systematize scientific information in a common vocabulary, suitable for exchanging pieces of information. Ontologies as such can be regarded as supportive means for scientific discovery processes, mainly because they facilitate reasoning and information analysis for a given problem domain (Joshi et al., 2004). Therefore, ontologies are mostly used for formally representing domain knowledge. They typically contain objects, concepts, properties and relations between objects.

The founder of the hypothesis formation field is Swanson (Swanson, 1986; Swanson et al., 2006). More that twenty years ago he demonstrated that new information could be generated by inspecting and relating two disjoint sets of articles from a given domain. Swanson showed that the sequence of causal reasoning from the facts published in medical literature could lead to forming hypotheses about causes of rare diseases. As an example he pinpointed pairs of articles that led him to conjecture a causal relation between migraine and magnesium. An example of such pair is the following: the first article stated that stress is related to migraine, while in the second article the authors claimed that stress could cause magnesium loss from the body. Starting from several such pairs of articles Swanson formed a hypothesis stating

that the depletion of magnesium can cause migraine attacks. Such hypothesis was not published yet in the literature up to that time; in the following months it was tested and proven correct. Even though the Swanson's approach requires substantial amount of manual work and is, therefore, only partially automatic, it was accepted as a promising one among several researchers that continued and upgraded his early work.
In this paper we first shortly describe approaches to ontology construction and give a practical illustration for the task. Then, we present and discuss a method for hypotheses generation that is based on Swanson's approach. Next, we demonstrate how ontology construction and text mining approach can be used to improve comprehension of written documents. We discuss possible applications of both methods and present our experience on medical and e-government fields.

2 Ontology construction

Traditionally, ontology construction is a manual task that consists of determining interesting domain concepts and establishing a hierarchy of such concepts. The process uses a special sort of description language, in which common domain knowledge is then represented. In the last decade several computer programs have been developed that support and speed–up such manual ontology construction, for example Protégé (Gennari et al., 2002). Since manual ontology construction is a complex and intricate process that requires both skill and diligence, the need for developing more active and helpful computerized support is evident.
With the emergence of new text mining technologies ontologies can be constructed semi-automatically by processing available text documents. In the last decade several approaches for facilitating semi-automatic construction of ontologies have been developed and successfully used in practice, making the process of ontology construction more effective and viable. One example of a tool for interactive construction of ontologies from text documents is OntoGen (Fortuna et al., 2006), which has already been proven successful in several real-world applications. A user can form concepts, edit them thematically and assign documents to the formed concepts. By implementing several modern machine learning techniques OntoGen helps users in all crucial phases of ontology construction, suggesting concepts and their names and automatically assigning documents to the proposed concepts (Fortuna, 2006).
An example of ontology constructed with OntoGen on the articles about autism is presented in Figure 1 (Cestnik et al., 2007). The articles were obtained from PubMed medical database, filtered so that only the articles published in the last ten years were taken as a result set for input. Figure 1 shows five main areas of the autism research: environmental factors, autism treatments, genetics, epidemiology, and neurobiology. The constructed top-level ontology gives an insight into the structure of the studied domain; it is, therefore, particularly useful in the process of obtaining initial acquaintance with the domain. Note that it provides some sort of "birds-eye view" on the domain under study. It is worth mentioning that the constructed ontology was meaningful and correct to the expert from the field of autism.
The importance of using ontologies has recently been demonstrated also in the field of e-government. Here, ontologies are used primarily to document e-government services and to establish a common ground for application integration and data exchange. E-government incorporates several services to citizens and legal entities. Typical e-government services for citizens include secure access to information and services about life events (e.g. marriage, childbirth, etc.). Services typically include various data types and models. Ontologies are used to facilitate easier and more reliable communication between e-government and citizens as well as between various sectors within e-government itself. In such was ontologies are key enablers of information interoperability between different applications and can have major impact on the success of e-government operations from the perspective of citizens.

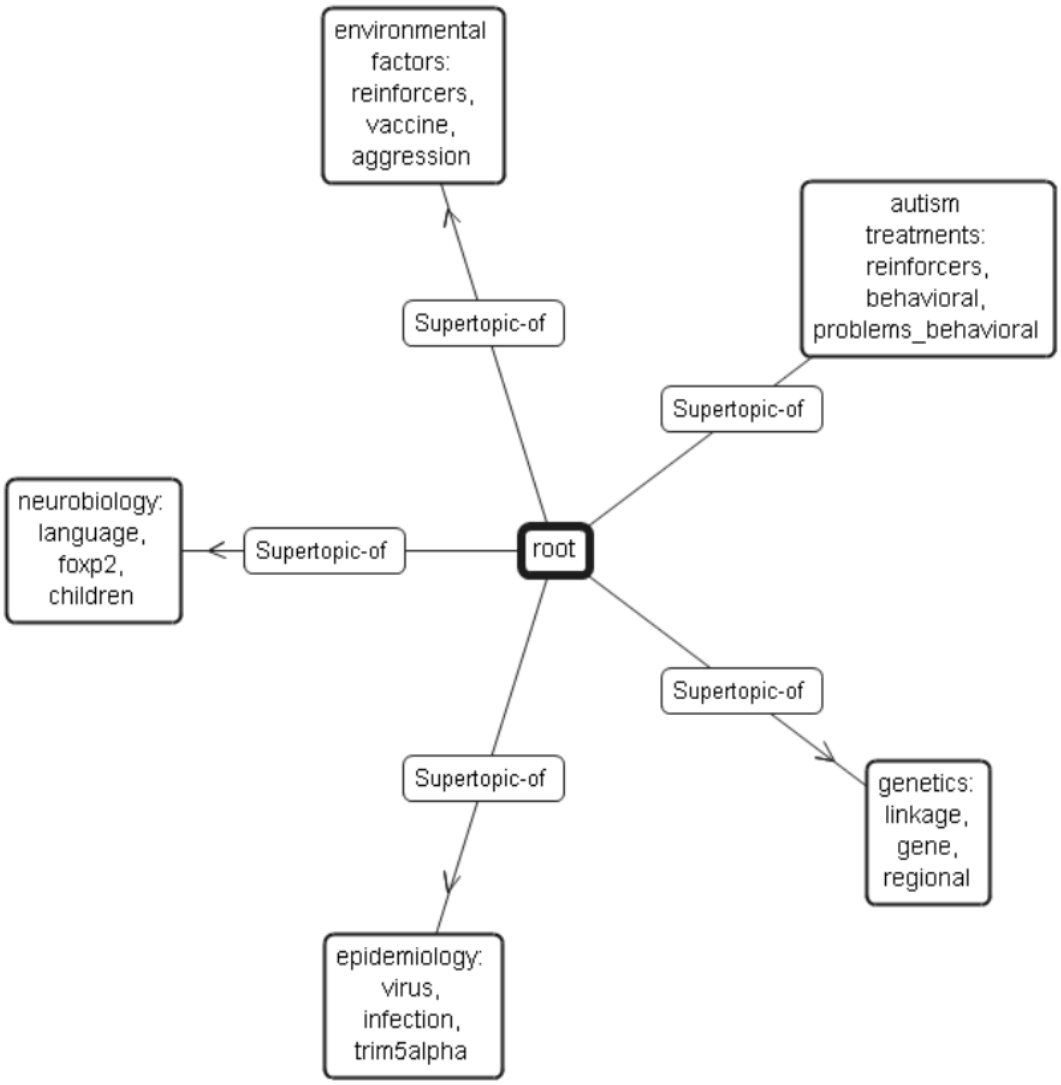

Figure 1: Semi-automatically constructed ontology from 217 articles about autism from medical database PubMed. The concept names were proposed by OntoGen and refined by the domain expert.

One of the e-Government projects that deal with knowledge management and ontologies is a EU project named QUALEG (Quality of Service and Legitimacy in e-Government) (Sagev et al., 2007). Its aim is to enable local governments of France, Poland and Germany to manage their policies in a transparent way and enable adaptability of the proposed solutions. The project showed many advantages of using technology based on ontology, such as preventing redundancy of data representation, enabling adaptability and supporting the realization of the importance ascribed by the local language to topics through the use of multiple synonyms.
Another EU funded project that deals with this context is the project LEX-IS. Its main objective is to improve the legislation process in the National Parliaments through enhancing public participation with the use of technology-based tools like ontologies. Developing ontologies helped all of the involved parties to easily locate and interpret important information (Loukis et al., 2007).

3 Hypothesis generation

In 1986 Swanson demonstrated that it is possible to discover hidden relations between biomedical concepts by inspecting published articles from a given domain (Swanson, 1986). Groups of articles where articles from one group do not cite articles from another group are referred to as disconnected literature. Swanson's main contribution was the discovery that such disconnected groups of articles could be linked together by using concepts that are common to all groups. When two or more such groups are linked by mutual concepts, the newly discovered connections represent potential source of new knowledge. For example, if literature *A* treats phenomenon *a*, which is supposedly related to phenomenon *b*, and literature *C* relates phenomenon *c* with phenomenon *b*, it can be assumed that literature *B* contains potential link between literatures *A* in *C*. According to the described method, Swanson reviewed the literature of that period and conjectured the hypothesis about causal relationship between the depletion of magnesium and some migraine types. The approach was popularly named ABC by the letters used to denote the groups of articles.

In The Center for systems and information technologies at The University of Nova Gorica a system RaJoLink has been developed that is designed to help experts generating and testing hypotheses within a given problem domain (Petrič et al., 2008). The system implements Swanson's ABC approach to hypotheses generation, promoting it to a new level of open discovery where discovering a target concept is a result of the process itself. RaJoLink implements a novel principle of rarity by searching for linking terms between literatures about rare technical terms that appear in the domain literature. Relations between such linking terms and the problem domain can be viewed as candidates for novel scientific hypotheses that are presented to the expert and further evaluated in the literature. Usually, RaJoLink is applied to text documents from various sources, assuming that each source represents a different context. The main motivation behind that task is to discover new knowledge by using the principle of bisociations. For an illustrative example, RaJoLink was applied in the autism domain. By analyzing scientific papers about autism we were able to discover hypothetical relations between autism and calcineurin and between autism and transcription factor NF-kappaB. Both hypotheses were medically acknowledged as an interesting contribution to the field of autism (Petrič et al., 2008; Urbančič et al., 2007). The autism expert was able to quickly search and evaluate her own hypotheses and, at the same time, semi-automatically generate new, potentially interesting hypotheses.

The size of the search space when exploring bisociative relations between different contexts is typically extremely large. Consequently, also the time complexity of the search is huge. Since the search method incorporates a domain expert to evaluate the most promising directions and steer the search, this additionally diminishes the effectiveness of the search for new knowledge, mostly due to the fact that for helpful guidance the expert has evaluate large set of possible options. The goal is to use the expert's knowledge to reduce the search space and thereby improving the efficiency of the developed method. By reducing the strain on the involved expert her efficiency grows up, consequently increasing the probability of achieving higher quality results with RaJoLink method. In turned out that filtering entities in relations with keywords from Medical Subject Heading (also referred to as MeSH filtering) can significantly contribute to the above task.

To facilitate easier collaboration with the domain expert the obtained results were presented in the form that is understandable and suitable for frequent expert's interpretations. Generated knowledge is represented in the form of knowledge graphs (Zhang, 2002), which constitute a firm basis for domain knowledge representation and knowledge exchange with other systems.

4 Improving comprehension of texts

In the modern western society citizens' lives are governed by excessive amount of regulations and laws. To function as a constitutive element of a society, one has to comply with all of them. However, besides the overly large quantity of legal regulations and laws, they are usually written using very complicated vocabulary and syntax, which in turn makes them hard to comprehend for an average citizen. One of the goals of e-government is to maintain high level of citizen satisfaction with its services. In this line, ontologies are sometimes used to describe and represent the requirements of specific regulations and laws to effectively convey the key concepts to target group of citizens.

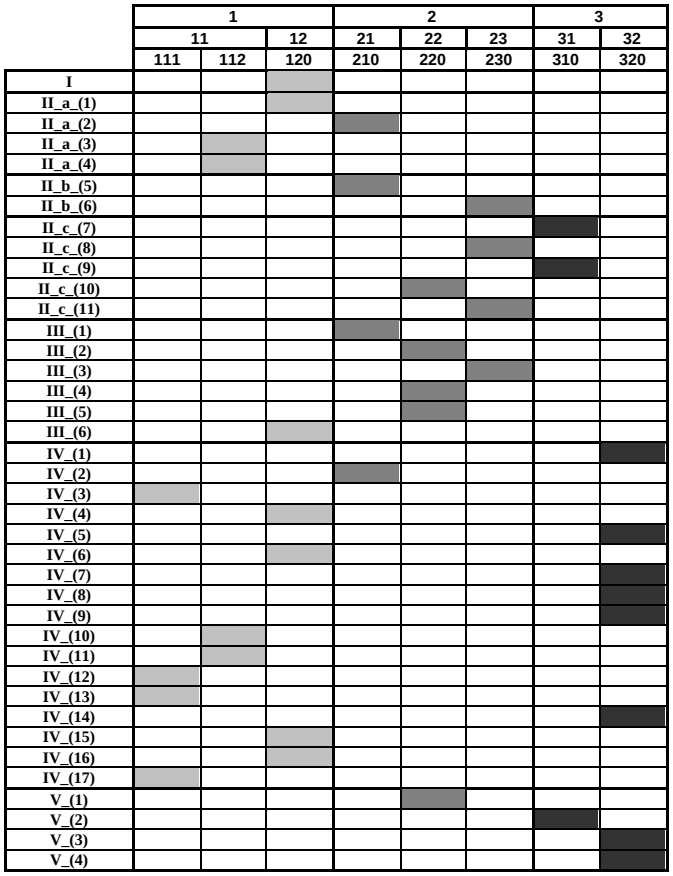

Figure 2: Labeled paragraphs of a selected legal document (lines) with their classification into ontology classes (columns).

Efficient and effective communication with the citizens is one of the top priorities of public administration. Citizens are the target population that consumes public administration services. Public tenders are one example of such services. Creating the text that includes requirements for a public tender is therefore demanding and time consuming task. However, it often happens that the final outcome is still unclear and hard to comprehend for the general public. One of the key issues here is the inherent complexity of the requirements. That is why the effort to improving the understandability is well justified.

Legal documents are in most cases similar to mathematical articles; within a given context they are supposed to cover and handle all the possibilities. Therefore, the loss of understandability is sometimes the price we have to pay for the sake of completeness. Even though the document contents might be syntactically and semantically correct, it can be

difficult to understand for non-expert public. Improving clarity of legal documents can reduce intricacies that occur due to misunderstanding or misinterpretation. Therefore, improving the comprehension is the goal of all public bodies that are responsible for preparing legal documents that are intended for use by the general public.

The method for improving the comprehension of legal documents (Cestnik et al., 2008) consists of several cycles. The goal of each cycle is to improve the structure and contents of outlier paragraphs. Outlier paragraphs are selected according to the visualization of the document by using high-level ontology concepts. The underlying assumption is that in a well-structured document that is easy to comprehend neighboring paragraphs treat similar topics using similar wording and expressions.

Each cycle of the proposed method consists of six phases: decomposition into paragraphs, text preprocessing, ontology construction, visualization, identification of outlier paragraphs, and correction of the identified outliers. An example of a result from one such cycle is illustrated in Figure 2. Starting from such visualization, paragraphs that need revising are identified. There are several actions that can be carried out to improve the comprehension of the document. First, such outlier paragraphs can be rewritten using more suitable wording. Second, they can be moved to more suitable context within the document. The new context is determined in such way that the neighboring paragraphs fall in similar ontology class as the identified paragraph. Last but not least, it can be decided and argued that the paragraph is clearly written and that its position within the document is well justifiable; in such way the experts can override the suggestion proposed by the method and OntoGen.

5 Conclusions

In the paper we presented the use of a tool OntoGen (Fortuna, 2006) for semi-automatic construction of ontologies on a collection of scientific articles from a given field. We demonstrated how such technology could help us answering questions like, for example, which are the areas of research in a given field, in which areas the research activity is the most intensive, how could the domain knowledge be structured into ontology. As a result, we can significantly speed-up the process of "getting acquainted" with a given problem domain, mostly because the generated birds-eye view quickly reveals the most important top-level domain concepts.

As one of the promising text mining results we described the use of RaJoLink method (Petrič et al., 2008) for discovering new scientific hypotheses in the field of autism. RaJoLink implements both closed discovery process, where hypotheses are known in advance, and open discovery process, where also tested hypotheses are generated within the process. The results of RaJoLink on the scientific articles about autism led to several important hypotheses that were recognized by the expert as an interesting insight towards better understanding of autism (Urbančič et al., 2007). By using RaJoLink, the expert was able to evaluate her own hypotheses more effectively (closed discovery process), as well as generate new, potentially interesting hypotheses (open discovery process).

We also presented the method for improving comprehension of legal documents that is based on semi-automatic ontology construction from a given document (Cestnik et al., 2008). The main advantage of using the proposed method lies in the fact that by improving the comprehension of legal documents we can improve the understanding of common rules and requirements for all the included stakeholders and actors. The method is applicable to any legal document for improving its understandability. Additional result of the method is the constructed ontology of top-level concepts that can be used to describe and present the corresponding legal document in more abstract way, as well as to visualize the document in a simple table.

Acknowledgement

The work about hypotheses generation in the field of autism presented in this paper was carried out at the University of Nova Gorica in the group with I. Petrič, T. Urbančič and M. Macedoni-Lukšič. The RaJoLink method was also developed in this group.

Literature

Cestnik B., Kern A., Modrijan H. (2008): Semi-automatic Ontology Construction for Improving Comprehension of Legal Documents. Electronic Government, 7th International Conference, EGOV 2008, Turin, Italy, August 31 - September 5, 2008. Proceedings, Lecture Notes in Computer Science, Springer, pp. 328-339.

Cestnik B., Petrič I, Urbančič T., Macedoni-Lukšič M. (2007): Structuring domain knowledge by semi-automatic ontology construction. *Organizacija (Kranj)*, 40(6), pp. 233-238.

Fortuna B., Grobelnik M., Mladenić D. (2006): System for semi-automatic ontology construction. Demo at ESWC 2006. Budva, Črna Gora, Junij, 2006.

Fortuna B. (2006): [http://ontogen.ijs.si/index.html], OntoGen: Description.

Gennari J., Musen M. A., Fergerson R. W., Grosso W. E., Crubezy M., Eriksson H., Noy N. F., Tu S. W. (2002): The Evolution of Protégé: An Environment for Knowledge-Based Systems Development.

Hutchins J. (1999): Retrospect and prospect in computer-based translation. Proceedings of MT Summit VII, pp. 30–44.

Joshi A., Undercoffer J.L. (2004): On Data Mining, Semantics, and Intrusion Detection. What to Dig for and Where to Find It. In: Data mining. Next Generation Challenges and Future Directions. Menlo Park, California. pp. 437-460.

Loukis E., Wimmer M.A., Charalabidis Y., Triantafillou A., Gatautis R. (2007): Argumentation systems and ontologies for enhancing public participation in the legislation process. Electronic Government, 6th International EGOV Conference, Regensburg, 3.-6.9.2007. Proceedings of ongoing research, project contributions and workshops, A. Groenlund, H.J. Scholl, M.A. Wimmer (Eds.), Trauner Verlag, Linz.

Petrič I., Urbančič T., Cestnik B., Macedoni-Lukšič M. (2008): Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics (2008), doi: 10.1016/j.jbi.2008.08.004.

Sagev A., Gal A. (2007): Putting Things in Context: A Topological Approach to Mapping Contexts to Ontologies, Journal on Data Semantics IX, str. 113-140.

Sebastiani F. (2002): Machine learning in automated text categorization. ACM Computing Surveys, 34 (1), pp. 1–47.

Singhal A. (2001): Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4), pp. 35–43.

Swanson D. R. (1986): Undiscovered public knowledge. Libr Q 1986, 56(2): 103-18.

Swanson D. R., Smalheiser N. R., Torvik V. I. (2006): Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). J Am Soc Inf Sci Tec 2006, 57(11): 1427-39.

Urbančič T., Petrič I., Cestnik B., Macedoni-Lukšič M. (2007): Literature mining: Towards better understanding of autism. Lect. notes cmput. sci., Artificial intelligence in medicine, pp. 217-226, Springer.