COMPARISON OF ONTOLOGIES BUILT ON TITLES, ABSTRACTS AND ENTIRE TEXTS OF ARTICLES

Ingrid Petrič[1], Tanja Urbančič[1,2], Bojan Cestnik[2,3]
[1]University of Nova Gorica
Vipavska 13, 5000 Nova Gorica, Slovenia
[2]Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
[3]Temida, d.o.o.
Dunajska 51, 1000 Ljubljana, Slovenia
e-mail: ingrid.petric@p-ng.si, tanja.urbancic@p-ng.si, bojan.cestnik@temida.si

ABSTRACT

This work investigates the differences in automatically constructed ontologies from titles, abstracts and bodies of texts, respectively. Articles about autism from database PubMed Central served as our testbed. In the comparison of autism onthologies, built with OntoGen, we focuse on vocabulary level of results of automatic concepts construction. A graphical presentation of comparison results is also proposed. The experiments show high similarity between ontologies built on abstracts and ontologies built on texts in the case of a subfield with specific terminology while in other cases, the role of whole texts was more important.

1 INTRODUCTION

Ontologies have the capability to share a common understanding of domains, and therefore to support a research with ability to reason over and to analyze the information at issue (Joshi, Undercoffer, 2004). Every ontology should efficiently communicate the intended meaning of analyzed context (Gruber, 1993). In the recent years, many tools that help constructing ontologies from texts in a given problem domain were developed and successfully used in practice (Brank et al, 2005). Among them, OntoGen (Fortuna et al, 2006) achieved a remarkable attention in the text-mining community.

Our comparison of ontologies, built with the tool OntoGen, was made on 214 articles from database PubMed Central that treat problems of autism. When dealing with complex phenomena, one good strategy is to decompose it to more manageable parts (Zupan et al., 1997). The document corpora can be usually divided by hierarchical structure of a document into logical sections such as title, abstract and main body (Hollingsworth, 2005). For these reasons we compared the different ontologies built with OntoGen on titles, abstracts and texts (main bodies) of PubMed articles, also to find out the most objective definitions of autism concepts. In addition, upon finding autism as a multilevel, complex phenomena, our goal was also to review the autism literature and to identify the most frequent topics researched in this domain.

Autism belongs to a group of pervasive developmental disorders, that are characterised by early delay and abnormal development of cognitive, communication and social interaction skills of a person. In the fourth edition of Diagnostic and Statistical Manual of Mental Disorders, revised, the category of pervasive developmental disorders refers to a group of symptoms of neurological development, connected with early brain mechanisms, that in large extent condition the social abilities already in the childhood (American Psychiatric Association, 2000). Heterogeneity of this developmental disturbance and its different degrees of affecting children has led to contemporary naming of autism with term: *Autism spectrum disorders*, to which suits the abbreviation *ASD*. Among data on autism there is often used the term *Asperger syndrome* together with the term *autism*. There are few content similarities between Asperger syndrome and autism, where no mental retardation is present (Klin, Volkmar, 1995). Both disorders are diagnostically placed within the group of autism spectrum disorders (American Psychiatric Association, 2000).

In this article we investigate the impact of how the inclusion and exclusion of various parts of scientific articles from the autism domain affect the constructed ontologies. More specifically, we study the differences in automatically constructed ontologies from titles, abstracts and bodies of texts respectively. First, we describe the origin and preparation of input texts. Then, we present the process of ontology construction with OntoGen. Next, we compare the obtained ontologies on vocabulary level and analyze the observed similarities and differences.

2 TARGET DATASET

For the purpose of mining the data on autism, we chose to analyze the professional literature that is publicly accessible on the World Wide Web in the data base of biomedical publications, PubMed. In the PubMed data

base there we found 10.821 documents (till August 21, 2006) that contain derived forms of *autis**, the expression root for autism. Between them there were 354 articles for which also their entire text has been published in the PubMed Central data base. Other relevant publications were either restricted to abstracts of documents or their entire texts were published in sources outside PubMed. From the listed 354 articles we further restricted the target set of articles on documents to those that have been published in the last ten years. As a result, we got 214 articles from 1997 forward, which we decomposed to titles, abstracts and texts for the analysis purpose.

3 ONTOLOGIES CONSTRUCTION BY ONTOGEN ON DOCUMENTS FROM PUBMED CENTRAL

OntoGen is a tool that enables interactive construction of ontologies about certain domain. The input for the tool is a collection of text documents. The user can create concepts, organize them into topic ontology and also assign documents to concepts. With the use of machine learning techniques OntoGen supports individual phases of ontology construction by suggesting concepts and their names, by defining relations between them, and by automatic assignment of documents to the concepts (Fortuna, 2005).
From the 214 documents, which had also their whole text published in the PubMed Central data base since 1997, the next 3 input text files were made: the file of 214 titles, the file of 214 abstracts, and the file of 214 texts (without their titles and abstracts). Each text file was entered into OntoGen, that from the entry data forms a model of most frequent terms and relations between them by K-means clustering technique. K-means algorithm tries to find such groups of documents that share similar words (Fortuna et al., 2006). The ontologies were built with two values for parameter k (for K-means algorithm in OntoGen): first, with parameter k=8, as automatically suggested by OntoGen, and second, with parameter k=5 that turned out to be a balanced tradeoff between the complexity and comprehensibility in this domain. In this way we got 8 and 5 topics respectively on the first level of domain ontology on entered titles, on abstracts and on entire texts of 214 autism documents. The concepts names (Keywords) and numbers of related documents (No. Docs) are presented by parts of OntoGen's screenshots in figures 1-6.

Id	Keywords	No. Docs
0	root	214
1	preference, assessment, effects	31
2	reinforcement, children_autism, early	27
3	genes, susceptibility, specific	32
4	functioning, syndrome, analysis	26
5	autism, teach, child	25
6	vaccination, schedules, activated	24
7	social, evidence, chromosome	17
8	disorders, linkage, case	32

Figure 1: *Concepts of autism ontology with 8 subgroups of 214 titles from the PubMed Central data base.*

Id	Keywords	No. Docs
0	root	214
1	sensory, sounds, auditory	8
2	stereotypy, behavioral, problems_behavioral	26
3	reinforcers, preferred, stimulus	41
4	teach, question, procedure	18
5	gene, linkage, regional	60
6	parent, mmr, vaccine	16
7	language, age, children	28
8	vaccine, mmr, mmr_vaccine	17

Figure 2: *Concepts of autism ontology with 8 subgroups of 214 abstracts from the PubMed Central data base.*

Id	Keywords	No. Docs
0	root	214
1	executive, nv, cortical	26
2	stereotypies, reinforcement, problems_behavior	27
3	reinforcement, session, aggression	38
4	prompted, script, teaching	21
5	linkage, family, gene	55
6	ht, secretin, legs	8
7	chemical, infant, sleep	14
8	vaccine, mmr, mmr_vaccine	25

Figure 3: *Concepts of autism ontology with 8 subgroups of 214 texts from the PubMed Central data base.*

Id	Keywords	No. Docs
0	root	214
22	autism, children_autism, children	67
23	syndrome, detection, social	19
24	disorders, spectrum, neurodevelopmental	39
25	genetic, chromosome, linkage	50
26	reinforcement, effects, behavior	39

Figure 4: *Concepts of autism ontology with 5 subgroups of 214 titles from the PubMed Central data base.*

Id	Keywords	No. Docs
0	root	214
22	reinforcers, behavioral, problems_behavioral	49
23	language, foxp2, children	52
24	reinforcers, vaccine, aggression	46
25	linkage, gene, regional	55
26	virus, infection, trim5alpha	12

Figure 5: *Concepts of autism ontology with 5 subgroups of 214 abstracts from the PubMed Central data base.*

Id	Keywords	No. Docs
0	root	214
22	reinforcement, session, trial	72
23	reinforcement, sleep, infant	37
24	vaccine, mmr, mmr_vaccine	24
25	linkage, family, gene	71
26	infection, pml, patients	10

Figure 6: *Concepts of autism ontology with 5 subgroups of 214 texts from the PubMed Central data base.*

The distribution of documents among 8 subgroups of titles ontology (Fig. 1) is quite uniform. Differently from that, the ontologies of 8 abstracts subgroups (Fig. 2), and 8 texts subgroups (Fig. 3), both show one major subgroup of documents that treat genetics, and another important group that describe reinforcers or stimulus for autists. Documents distributions in ontologies of 5 subgroups are a little different. There are two major groups of titles (Fig. 4) and texts (Fig. 6). The biggest group of titles describe autism in general, whereas the largest texts group writes about reinforcement trials. The second major group in both cases (titles and texts) deals with genetics. Abstracts distributions (Fig. 5), on the contrary show two most important groups that both treat genetics in some way. The first one has clear genetic keywords. The second group includes, beside others, the keyword *foxp2*, which is a gene important for the development of speech.

4 COMPARISON OF AUTISM ONTOLOGIES ON VOCABULARY LEVEL

Ontologies are complex structures, where there is often more reasonable to focus the attention on the evaluation of separate levels of ontology, rather than on the direct evaluation of whole ontology (Brank et al., 2005). In our comparison of autism ontologies, built with OntoGen, we focused on vocabulary level of results of automatic concepts construction about autism domain. We observed distribution of documents within individual ontology groups on the first level of each ontology model (first level subgroups of autism domain), considering terminology that was chosen by OntoGen for presentation of concepts. In addition, with the support of OntoGen, we tried to evaluate also the content compliance of titles and summaries to belonging entire texts of related documents.

Clustering algorithms, such as K-means, are useful tools for data mining; however when we have to cluster datasets, it is not always clear, which is the most suitable number of clusters (parameter *k*) to use (Hamerly, Elkan, 2003). OntoGen automatically proposes 8 clusters, beside this the user should try with different *k*-s in order to find out the best result for the investigated domain.

4.1 Comparison of ontologies, analysed by parameter *k*=8 in K-means clustering

The evaluation of different results, next to changing the parameter *k* when analysing the phenomenon of autism as it is described in 214 documents from PubMed Central data base, showed differences between conceptual design of titles, abstracts and related texts. Our graphic display in figure 7 is result of comparison of ontologies above abstracts and texts, that were analysed by parameter *k*=8 in K-means clustering performed by OntoGen. From the comparison between ontology of 8 texts groups versus 8 abstracts groups, the major similarity is shown between the group of genetic documents, which include the same 40 articles from the observed dataset. An important similarity is seen also between the group of texts and the group of articles, that talk about reinforcement. Without the specific similarity with groups of abstracts remains only the smallest group of texts, with keywords: *ht, secretin, legs*. From the keywords of this group and by the contextual knowledge of the autism phenomenon we deduce, that in this case, the group is related to documents which present the concepts, that are rarely mentioned in autism context.

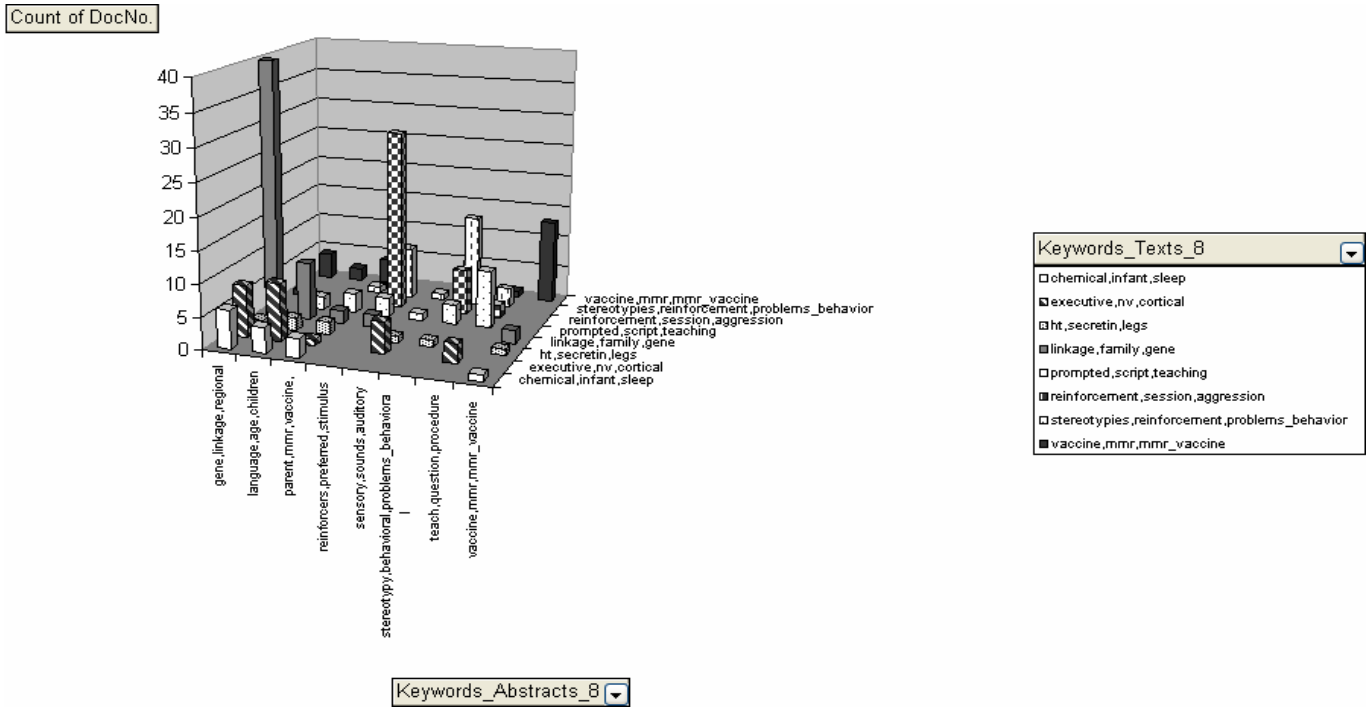

Figure 7: *Comparison between the distribution of abstracts and texts in ontologies with 8 subgroups.*

Compared to analysis of texts and abstracts matching, there is significantly lower similarity between texts and titles or between abstracts and titles of relative articles. Articles about genetics are the only rather important group of documents that have some more similar vocabulary both in texts versus titles, as well as in abstracts versus titles comparison. This is due to the genetic terminology, and to the genetic context itself, which is very specific, if compared to other researching fields of autism.

4.2 Comparison of ontologies, generated with parameter *k*=5

After having analysed the autism domain by parameter *k*=8, we analysed autism, as it is presented by 214 documents from PubMed Central data base, by *k*=5 and many other different parametres as well. Among the groups of documents which belong to the certain of 5 subgroups of texts and at the same time to its relative subgroup of abstracts, the largest similarity is between the group of genetic texts and abstracts, which cover the same 51 documents. Relatively large similarity is seen also between the texts and the abstracts groups that deal with virus infections. Less similarity is between the group of texts and corresponding abstracts subgroup about MMR vaccine. Even less specific is similarity between abstracts and texts from groups: *reinforcement, session, trial* and *reinforcement, sleep, infant*. In this case we can notice one of the keywords used twice, as a part of definition of two separate concepts in texts ontology (the term *reinforcement*), like in abstracts ontology (the term *reinforcers*).

The comparison of ontologies with 5 groups of texts and 5 groups of titles shows the biggest similarity between the groups of texts and titles on genetics, as well as between the group of texts: *reinforcement, session, trial* and a group of titles, to which belong keywords: *autism, children_autism, children*. Besides the already mentioned genetics articles, there are no specific similarities between the ontology of abstracts and the ontology of titles.

5 CONCLUSIONS AND FUTURE WORK

When trying to identify some useful knowledge from huge volumes of digital data, rather then reading and manually analysing all available data, which is a time consuming task, we can guide our attention only on the most relevant information above domain of interest. Such information can be identified by ontologies construction, that we found as a very fast and effective way of exploration of large datasets. Ontologies actually helped us to review and understand the complex and heterogeneous specter of scientific articles about autism.

Comparison of ontologies is, such as ontology itself, complex and requires thorough examination of possible causes for revealed distributions of documents inside certain ontology. Our graphic presentation of compared ontologies clearly exposes the main clusters of autism articles, which are shown as the highest columns in the graph. Thus it provides a powerful way to visualize the biggest similarities in observed ontologies, where we can see that the largest collection of autism documents always deal with genetics. The only exception comes from the comparison of ontologies of 5 texts groups and 5 titles groups, that beside genetics, gives importance also to the keyword *autism* (*autism, children_autism, children*). Here rises the question, whether the distribution of documents within ontologies, woud be the same, if we delete from the input files those expressions, that mark the domain itself, as is in our case the term *autism* and its derivatives. In our opinion, this entry data preprocessing step could significantly change distributions of documents between groups of ontology and would contribute to the clearer identification of ontology concepts.

Acknowledgement

This work was partially supported by the Slovenian Research Agency programme Knowledge Technologies (2004-2008).
We thank Nada Lavrač for her suggestion to use OntoGen and Blaž Fortuna for his discussions about OntoGen's performance. We also appreciate help and support we got from Marta Macedoni-Lukšič in our efforts to better understand autism.

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision. Washington, DC, 2000.
2. Brank J, Grobelnik M, Mladenić D: A survey of ontology evaluation techniques. In: SIKDD 2005 at multiconference IS 2005, Ljubljana, Slovenia, 17 October 2005.
3. Fortuna B:[http://ontogen.ijs.si/index.html], OntoGen: Description, 2006.
4. Fortuna B, Grobelnik M, Mladenić D: System for semi-automatic ontology construction. Demo at ESWC 2006. Budva, Montenegro, June, 2006.
5. Gruber TR: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In: Formal Ontology in Conceptual Analysis and Knowledge Representation. Padova, Italy, 1993.
6. Hamerly G, Elkan C: Learning the k in k-means, Proc. of the Neural Information Processing Systems 2003, [http://citeseer.ist.psu.edu/hamerly03learning.html].
7. Hollingsworth B, Lewin I, Tidhar D.: Retrieving Hierarchical Text Structure from Typeset Scientific Articles - a Prerequisite for E-Science Text Mining. In: Proceedings of the 4th UK E-Science All Hands Meeting, Nottingham 2005, 267-273.
8. Joshi A, Undercoffer JL: 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, 2004. 437-460.
9. Klin A, Volkmar FR: Asperger's Syndrome, Guidelines for Assessment and Diagnosis. Pittsburgh, Learning Disabilities Association of America, June 1995.
10. PubMed Central [http://www.ncbi.nlm.nih.gov/]
11. Zupan B, Bohanec M, Bratko I, Cestnik B: A Dataset Decomposition Approach to Data Mining and Machine Discovery. In: KDD 1997, Ljubljana, Slovenia. 299-302.