Discovering Hidden Knowledge from Biomedical Literature

Ingrid Petrič[1], Tanja Urbančič[1,2] and 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

Keywords: text mining, ontology construction, autism

Received: June 3, 2006

In this paper we investigate the potential of text mining for discovering implicit knowledge in biomedical literature. Based on Swanson's suggestion for hypotheses generation we tried to identify potential contributions to a better understanding of autism focusing on articles from database PubMed Central. First, we used them for ontology construction in order to obtain an improved insight into the domain structure. Next, we extracted a few rare terms that could potentially lead to new knowledge discovery for the explanation of the autism phenomena. We present a concrete example of such constructed knowledge about a substance calcineurin and its potential relations with other already published indications of autism.

Povzetek: Prispevek opisuje uporabo metod rudarjenja besedil na medicinskih člankih s področja avtizma.

1 Introduction

The practice of biomedicine is, as well as other activities of our society, inherently an information-management task (Shortliffe, 1993). Internet, the very common and increasingly used information source, provides massive heterogeneous collections of data. Huge bibliographic databases thus often contain interesting information that may be inexplicit or even hidden. One of such databases is MEDLINE, the primary component of PubMed, which is the United States National Library of Medicine's bibliographic database. It covers over 4.800 journals published in more than 70 countries worldwide and thus contains over 14 million citations from 1966 to the present (PubMed, 2006). The daily increasing number of biomedical articles provides a huge potential source of new data. In MEDLINE database there are between 1.500-3.500 complete references added since 2002 each day from Tuesday to Saturday (PubMed, 2006).

There is an urgent need to assist researchers in extracting knowledge from the rapidly growing volumes of databases in order to improve the usefulness of these vast amounts of data. For such reasons, the ability to extract the right information of interest remains the subject of the growing field of knowledge discovery in databases. Knowledge discovery is the process of discovering useful knowledge from data, which includes data mining as the application of specific algorithms for extracting patterns from data (Fayyad et al., 1996). In fact, important information hidden in huge databases could be discovered by data mining and knowledge discovery techniques. More specifically, those databases that contain bibliographic semi-structured data can be approached by text mining as specific kind of data mining.

Although the technology for data and text mining is well advanced, its potential still seems to lack sufficient recognition. Healthcare in general is one of the slowest sectors in utilizing information and communication technologies to their full benefit; however, the need for computer literacy has already been recognised and acknowledged by professionals in this sector (Štepankova, Engova, 2006). Therefore, one of the major challenges of biomedical text mining over the next 5 to 10 years is to make these techniques better understood and more useful to biomedical researchers (Cohen, Hersh, 2005). At the same time, the continued cooperation with professional communities such as the biomedical research community is required to ensure that their needs are properly addressed. Such collaboration is particularly crucial in complex scientific areas, as for example in autism field of biomedical research. The specific requirements in autism research, as presented by Zerhouni (2004), actually emphasise the need for

increasing the efficiency of communication of research findings to the related science community.

Autism belongs to a group of pervasive developmental disorders that in most cases have unclear origin. The main characteristic components of abnormal functioning in autism are the early delay and abnormal development of cognitive, communication and social interaction skills of affected individuals. In the fourth, revised edition of Diagnostic and Statistical Manual of Mental Disorders, a 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). Such heterogeneous features of autistic developmental disturbance and its different degrees of affecting children have led to contemporary naming of autism conditions with the term: *Autism spectrum disorders*, to which suits the abbreviation *ASD*. The lack of studies, evidenced by Zerhouni (2004), that would increase the knowledge about risk factors and early development of autism, and that would better define characterization of autism spectrum disorders, has led us to choose the autism as an application domain of our research in knowledge technologies.

In this article we focus on the areas and methods where text mining potentially enriches biomedical science and thus interdisciplinary connects information technologies with biomedical expert knowledge. First we describe several text mining approaches in real biomedical settings towards extracting knowledge from data. Then we present our approach towards integration of real problem analysis and extraction of potentially useful information from data. Our main aim was to extract some implicit and previously unknown interesting information from professional articles about autism. Some of our text mining results are finally described with example pairs of implicit connections that we managed to identify from biomedical articles.

2 Text mining in biomedicine

There are several biomedical examples, where data mining has been successfully applied, as described in a review by Van Someren and Urbančič (2006). Examples include diagnosis, where data mining relates symptoms and other attributes of patients to their disease, subgroups of patients that are at risk for certain disease, and gene expression, with a growing number of applications, where predictions and identifications of disease markers are made, based on features of genes.

While data mining usually operates with collections of well structured data, researchers often have to deal with semi-structured text collections, too. Such datasets require the use of text mining techniques. Extracting important information from the increasingly available biomedical knowledge represented in digital text forms, has been proved as an important opportunity for biomedical discoveries and hypothesis generation. Having access and ability to work with the newest information, indeed means great potential for experts, who can benefit from the advantages of information systems and technologies. Biomedical informatics thus presents an essential element of biomedical research process. Methods that have been recently used for biomedical text mining tasks include the following items (Cohen, Hersh, 2005):

- *Named entity recognition* in order to identify all of the instances of a name for specific type of domain, within a collection of text;
  Examples of recent areas of biomedical research:
  - drug names within published journal articles,
  - gene names and their symbols within a collection of MEDLINE abstracts.

  Text mining approaches: lexicon-based, rules-based, statistically based, combined.
- *Text classification* with the goal to automatically determine whether a document or a part of it has particular attributes of interest;
  Examples of recent areas of biomedical research:
  - documents discussing a given topic,
  - texts containing a certain type of information.

  Text mining approaches: classification rule induction.
- *Synonym and abbreviation extraction* with the attempt to speed up literature search with automatic collections of synonyms and abbreviations for entities;
  Examples of recent areas of biomedical research:
  - gene name synonyms,
  - biomedical term abbreviations.

  Text mining approaches: combination of named entity recognition system, with statistical, support vector machine classifier-based, and automatic or manual pattern-based matching rules algorithms.
- *Relationship extraction* with the goal to recognize occurrences of a pre-specified type of relationship between a pair of entities of specific types;
  Examples of recent areas of biomedical research:
  - relationships between genes and proteins,
  - text-based gene clustering.

  Text mining approaches: neighbour divergence analysis, vector space approach and k-medoids clustering algorithm, fuzzy set theory on co-occurring dataset records, type and part-of-speech tagging.
- *Integration frameworks* with intention to address many different user needs;
  Examples of recent areas of biomedical research:
  - comparison of gene names and functional terms,
  - gene based text clusters.

  Text mining approaches: template-based, text profiling and clustering based.
- *Hypothesis generation* that focuses on the uncovering of implicit relationships, worthy of further investigation, that are inferred by the presence of other more explicit information;
  Examples of recent areas of biomedical research:

- connection between patient benefit and food substances,
- potential new uses and therapeutic effects of drugs.

Text mining approaches: Swanson's ABC model-based.

In the continuation we concentrate on hypotheses generation as a central point of our research interest.

3 Related work

The machine learning process is characterized by the search space, which reflects the expression of the hypothesis language, as a target knowledge (Botta et al., 2003). Idea of the text mining approach towards hypothesis generation, known as Swanson's ABC model, consists of discovering complementary structures in disjoint journal articles. This model assumes that when one literature reports that agent A causes phenomenon B, and second literature reports that B influences C, we could propose that agent A might influence phenomenon C (Swanson, 1990). To find some published evidence leading to undiscovered knowledge, the A and C literatures should have few or no published articles in common. In such way, Swanson discovered, among other, several relationships that connected migraine and decreased levels of magnesium (Swanson, 1990).

To facilitate the discovery of hypotheses by linking findings across literature, Swanson and his colleagues designed a set of interactive software that is available on a web-based system called Arrowsmith (Smalheiser, Swanson, 1998). Pratt and Yetisgen-Yildiz (2003) designed LitLinker that uses data mining techniques to identify correlations among concepts and then uses those correlations for discovery of potential causal links between biomedical terms. Sehgal et al. presented a system that may be used to explore topics and their relationships using text collections such as MEDLINE (Sehgal et al., 2003). Weeber et al. experimented with Swanson's idea of searching the literature for generating new potential therapeutic uses of the drug thalidomide with the use of a concept-based discovery support system DAD on the scientific literature (Weeber et al., 2003). Another example of discovering new relations from bibliographic database according to Swanson's model is identification of disease candidate genes by an interactive discovery support system for biomedicine Bitola (Hristovski et al., 2005). Transitive text mining was explored also by Grohmann and Stegmann (2005), who developed a web-based tool, C-MLink.

For successful data mining a wide background knowledge concerning the problem domain presents a substantial advantage. In fact, hypothesis generation from text mining results relies on background knowledge, experience, and intuition (Srinivasan, 2004). With this consideration we started our examination of autism phenomena with the identification of its main concepts and the review of what is already known about autism. We identified such information by ontologies construction, which we found a very fast and effective way of exploring large datasets. Ontologies in general with their capability to share a common understanding of domains support researches with ability to reason over and to analyse the information at issue (Joshi, Undercoffer, 2004). Many tools that help constructing ontologies from texts were developed and successfully used in practice (Brank et al., 2005). Among them, OntoGen (Fortuna et al., 2006), the interactive tool for semi-automatic construction of ontologies, received a remarkable attention.

4 Identification of domain structure

An important goal in our recognition of autism phenomena was to uncover the fundamental concepts that provide the common knowledge about autism. To identify some useful pieces of knowledge from the large amount of digital articles one approach would be to read and manually analyse all available data. Since this is evidently a time consuming task, we instead chose to guide our attention only on the most relevant information about the domain of interest. We performed our research with the computational support of OntoGen.

4.1 Target dataset

We decided to analyse the professional literature about autism that is publicly accessible on the World Wide Web in the database of biomedical publications, PubMed. In the PubMed database we found 10.821 documents (till August 21, 2006) that contain derived forms of autis*, the expression root for autism. There were 354 articles with their entire text published in the PubMed Central database. 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 purpose of further analysis.

4.2 Text mining support system

One of the most frequently used text representations in text mining is word-vector representation, where the word-vector contains some weight for each word of text, proportional to the number of its occurrences in the text (Mladenić, 2006). Such representations are used also by OntoGen, which enables interactive construction of ontologies in a chosen domain. We used it to construct several autism ontologies. The input for the tool is a collection of text documents. With machine learning techniques OntoGen supports important 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).

4.3 Ontology of autism domain

Our aim was first to review the autism literature and to identify the most frequent topics researched in this domain. With this intention we built the autism domain ontology with OntoGen on 214 articles from PubMed Central database that treat problems of autism. OntoGen displayed sub-concepts of autism domain as suggested by its clustering algorithm, and described them with their main keywords extracted from text documents. The keywords that we used for concepts description were calculated both according to the concept centroid vector, and by the Support Vector Machine based linear model. The system also displayed the current coverage of each concept by the number of documents that it positively classified into the concept and the inner-cluster similarity measures. Ontologies built with OntoGen, as an example shown in Figure 1, actually helped us to substantially speed up the process of reviewing and understanding the complex and heterogeneous spectrum of scientific articles about autism.

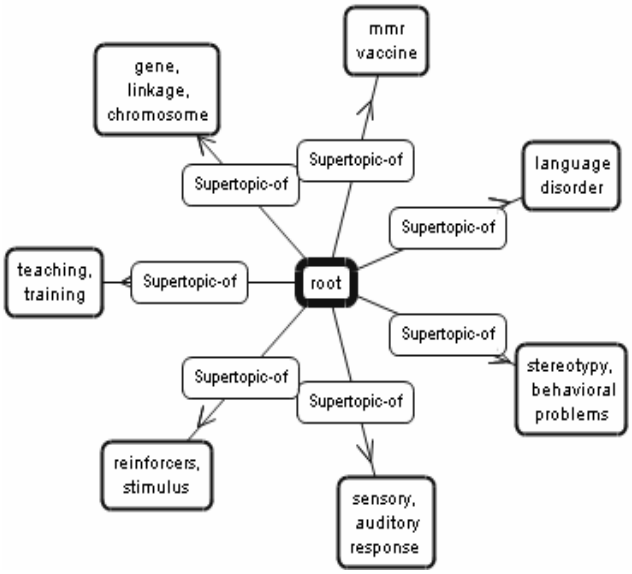

Figure 1: Concepts of autism ontology with 7 subgroups, built on 214 abstracts from the PubMed Central database.

The main concepts of autism phenomena as they result from the first level of our ontology model (first level subgroups of autism domain) are: genetics; teaching and training; reinforcers and stimulus; sensory and auditory response; stereotypy and behavioural problems; language disorders, and MMR (Measles, Mumps, and Rubella) vaccine. Important confirmation of the resulted ontology construction is the recent state of autism research as described by Zerhouni (2004) that summarizes the main scientific activities of autism research in the major areas of epidemiology, genetics, neurobiology, environmental factors and specific treatments of autism.

5 Extraction of implicit relationships from autism data

Besides constructing an ontology on the input file of texts, OntoGen creates also a *.txt.stat file with statistical incidence of terms as they appear in documents collected in the input dataset. We utilized this OntoGen's by-product as the basis for our approach toward the identification of rare relations between autism data. As our goal was to discover undocumented knowledge about autism phenomena, we assumed that starting our search on rare connections between data rather than on frequent ones, we would have better chances to discover implicit relations that are still unknown and might, however, be useful for the autism researchers.

5.1 The related concepts

Our approach towards discovering knowledge about autism concentrated on identifying interesting concepts within autism sub-areas of interest. Therefore, we considered the subdivision of autism domain on research fields; moreover, we particularly guided our attention on neurobiological basis of autistic abnormalities.

To find some related concepts, which would lead us to potential discoveries of new knowledge, we took the *.txt.stat file created by OntoGen while constructing ontologies. We first focused our attention on those terms listed in this text file that appeared only in one article from the input dataset. Taking into account also background knowledge about autism, we chose words that could be useful for autism discovery. Three of the chosen terms, presented also in the intersection area in Figure 2, are: lactoylglutathione, synaptophysin and calcium_channels. There are three major reasons for these choices. First, we found that an increase in polarity of glyoxalase I in autism brains was reported and that glyoxalase system involves also lactoylglutathione. Second, as the altered synaptic function was also discussed in autism articles, we took in consideration synaptophysin, a protein localized to synaptic vesicles. And third, abnormal calcium signalling was found in some autistic children, thus we chose also term calcium_channels for further discovery. After selecting these three terms of interest, we searched the article database to find what all these terms have in common.

One of the goals of text mining is to automatically discover interesting hypotheses from a potentially useful text collection (Srinivasan, 2004). By text mining on PubMed articles that treat these selected terms domains, we constructed their ontologies and from the OntoGen's *.txt.stat files we retrieved the words they all have in common (the words that appeared in the three *.txt.stat files). One of such terms, listed also in Figure 2, that could be interesting for the hypothesis generation and forward research on autism phenomena, is calcineurin. Calcineurin is calcium- and calmodulin-dependent serine/threonine protein phosphatase, which is widely present in mammalian tissues, with the highest levels found in brain (Rusnak, Mertz, 2000). Our literature mining in disjoint journal articles showed that it could be

related to autistic disorders, however to the present no direct evidence of calcineurin role in autism has been reported yet on the internet.

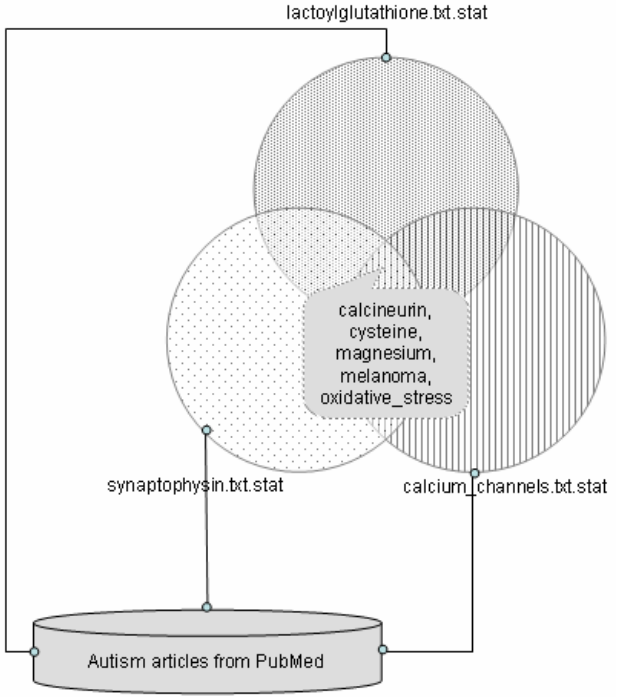

Figure 2: Results of our approach to literature mining on autism domain.

5.2 The explored conjectures

In order to justify the role of calcineurin in autistic problem domain we decided to search and explore possible reasoning paths that relate the selected substance to some known expressions of autism. Since the direct relation was not yet noted in the literature, our goal was to find a few plausible interconnecting terms that relate the two notions (Swanson, 1990). Having this in mind, we explored the union of PubMed articles about autism and articles about calcineurin. By building ontologies on such input dataset of combined articles the goal was to discover documents having as much as possible words in common. For this purpose we searched for the highest similarity measures inside the clusters of ontologies. Interestingly, by this search we identified several pairs of instances of PubMed articles that are connecting the two categories of biomedical literature, autism and calcineurin, respectively. This way we were able to find eleven pairs of articles, which, when put together, could be seen as arguments for new hypotheses of autism and calcineurin relationship, such as the three listed in Table 1.

When showing the presented results to the expert of autistic spectrum and related disorders, she not only confirmed strong interest in the method and in the discovered relations, but was also able to guide our further work very efficiently by turning our attention on discovering the relationship between autism and fragile X chromosome.

Autism literature	Calcineurin literature
Fatemi et al. (2001) reported a reduction of Bcl-2 (a regulatory protein for control of programmed brain cell death) levels in autistic cerebellum.	Erin et al. (2003) observed that calcineurin occurred as a complex with Bcl-2 in various regions of rat and mouse brain.
Qiu et al. (2006) described the low-density lipoprotein receptors that regulate cholesterol transport, in neuropsychiatric disorders, such as autism.	Cofan et al. (2005) published their article about effect of calcineurin inhibitors on low-density lipoprotein oxidation.
Bear et al. (2004) reported about the loss of fragile X protein, an identified cause of autism that increased long-term depression in mouse hippocampus.	Zhabotinsky et al. (2006) described induction of long-term depression that depends on calcineurin.

Table 1: Hypotheses for calcineuring and autism relationship.

6 Conclusion

Our study confirms the potential of ontology construction by OntoGen on biomedical literature to systematically structure main concepts. The evaluation of the ontology constructed on autism showed important similarity to the reported state of autism research.

Considering OntoGen's statistical data can lead to discovery of potentially useful and previously unknown information related to the researched phenomena. In such way, OntoGen's functionality can be extended to retrieve new information from vast amounts of textual data that experts otherwise have to explore manually. As connecting sets of literature about synaptophysin, lactoylglutathione and calcium channels that were selected as three interesting rare terms from autism articles, we found calcineurin, cysteine, magnesium, melanoma, oxidative stress and many others. In the preliminary expert evaluation the approach proposed in this paper proved to be successful. However, further assessment of the possible role of calcineurin and other resulting candidates in autism is needed to justify our methodological approach and to see if it can contribute to the knowledge corpus of autism phenomena.

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 (2000) Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision. Washington, DC.

[2] Bear M.F., Huber K.M., Warren S.T. (2004) The mGluR theory of fragile X mental retardation, *Trends in Neurosciences*, 27(7), pp. 370-377.

[3] Botta M., Saitta L., Sebag M. (2003) Relational Learning as Search in a Critical Region, *Journal of Machine Learning Research*, 4, pp. 431-463.

[4] Brank J., Grobelnik M., Mladenić D. (2005) A survey of ontology evaluation techniques, *SIKDD 2005 at multiconference IS 2005*, Ljubljana, Slovenia.

[5] Cofan F., Cofan M., Campos B., Guerra R., Campistol J.M., Oppenheimer F. (2005) Effect of calcineurin inhibitors on low-density lipoprotein oxidation, *Transplantation Proceedings*, 37(9), pp. 3791-3793.

[6] Cohen A.M., Hersh W.R. (2005) A Survey of Current Work in Biomedical Text Mining, *Briefings in Bioinformatics*, 6(1), pp. 57-71.

[7] Erin N., Bronson S.K., Billingsley M.L. (2003) Calcium-dependent interaction of calcineurin with Bcl-2 in neuronal tissue, *Neuroscience*, 117(3), pp. 541-555.

[8] Fatemi S.H., Stary J.M., Halt A.R., Realmuto G.R. (2001) Dysregulation of Reelin and Bcl-2 proteins in autistic cerebellum, *Journal of Autism and Developmental Disorders*, 31(6), pp. 529-535.

[9] Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996) Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon.

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

[11] Fortuna B., Grobelnik M., Mladenić D. (2006) System for semi-automatic ontology construction. Demo at ESWC 2006, Budva, Montenegro.

[12] Grohmann G., Stegmann J., C-MLink: a web-based tool for transitive text mining, *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Sweden, Stockholm, pp. 658-659

[13] Hristovski D., Peterlin B., Mitchell J.A., Humphrey S.M. (2005) Using literature-based discovery to identify disease candidate genes, *International Journal of Medical Informatics*, 74, pp. 289-298.

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

[15] Mladenić D. (2006) Text Mining: Machine Learning on Documents, *Encyclopedia of Data Warehousing and Mining*, Hershey: Idea Group Reference, pp. 1109-1112.

[16] Pratt W., Yetisgen-Yildiz M. (2003) LitLinker: Capturing Connections across the Biomedical Literature, *Proceedings of the International Conference on Knowledge Capture (K-Cap'03), Florida*, pp. 105–112.

[17] PubMed (2006) [http://www.ncbi.nlm.nih.gov/], Overview.

[18] Qiu S., Korwek K.M., Weeber E.J. (2006) A fresh look at an ancient receptor family: emerging roles for density lipoprotein receptors in synaptic plasticity and memory formation, *Neurobiology of Learning and Memory*, 85(1), pp. 16-29.

[19] Rusnak F., Mertz P. (2000) Calcineurin: Form and Function, *Physiological Reviews*, 80(4), pp. 1483-1521.

[20] Sehgal A., Qiu X.Y., Srinivasan P. (2003) Mining MEDLINE Metadata to Explore Genes and their Connections, *Proceedings of the SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics*.

[21] Shortliffe E.H. (1993) The Adolescence of AI in Medicine: Will the Field Come of Age in the '90s? *Artificial Intelligence in Medicine*, 5(2), pp. 93-106.

[22] Smalheiser N.R., Swanson D.R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses, *Computer Methods and Programs in Biomedicine*, 57, pp. 149-153.

[23] Srinivasan P. (2004) Text mining: Generating hypotheses from MEDLINE, *Journal of the American Society for Information Science and Technology*, 55(5), pp. 396-413.

[24] Swanson D.R. (1990) Medical literature as a potential source of new knowledge, *Bulletin of the Medical Library Association*, 78(1), pp. 29-37.

[25] Štepankova O., Engova D. (2006) Professional Competence and Computer Literacy in e-age, Focus on Healthcare, *Methods of Information in Medicine*; 45, pp. 300-305.

[26] Van Someren M., Urbančič T. (2006) Applications of machine learning: matching problems to tasks and methods, *The Knowledge Engineering Review*, 20(4), pp. 363-402.

[27] Weeber M., Vos R., Klein H., De Jong-van den Berg L.T., Aronson A.R., Molema G. (2003) Generating Hypotheses by Discovering Implicit Associations in the Literature: A case Report of a Search for New Potential Therapeutic Uses for Thalidomide, *Journal of the American Medical Informatics Association*, 10(3), pp. 252-259.

[28] Zerhouni E.A. for National Institutes of Health and National Institute of Mental Health (2004) Congressional Appropriations Committee Report on the State of Autism Research. Department of Health and Human Service, Bethesda, Maryland.

[29] Zhabotinsky A.M., Camp R.N., Epstein I.R., Lisman J.E. (2006) Role of the neurogranin concentrated in spines in the induction of long-term potentiation, *Journal of Neuroscience*, 26(28), pp. 7337-7347.