Ingrid Petrič

TEXT MINING FOR DISCOVERING IMPLICIT RELATIONSHIPS IN BIOMEDICAL LITERATURE

Doctoral Dissertation

TEKSTOVNO RUDARJENJE ZA ODKRIVANJE IMPLICITNIH POVEZAV V BIOMEDICINSKI LITERATURI

Doktorska disertacija

Supervisor: prof. dr. Tanja Urbančič

Co-Supervisor: doc. dr. Bojan Cestnik

September 2009

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenia

Index

Abstract ..... **VII**

Povzetek ..... **IX**

Abbreviations ..... **XI**

1 Introduction **1**

2 Aims and Hypothesis **3**
2.1 Motivation 3
2.2 Research hypotheses 4
2.3 Scientific contributions 5

3 Related Work **7**
3.1 Text mining and knowledge discovery 7
3.2 Link analysis 11
3.3 Creativity through bisociations 12
3.4 Swanson's model of knowledge discovery 13
3.5 Approaches based on Swanson's ABC model 14

4 Autism Domain **17**
4.1 Trend analysis in autism research 17
4.2 Recent autism research 19

5 Materials and Methods of Data Collection **21**
5.1 Research outline 21
5.2 Corpus collection 22
5.3 Records in XML format 23
5.4 MeSH classification 25

6 Structuring Domain Knowledge with Ontology Construction **27**
6.1 Ontologies 27
6.2 Semi-automatic ontology construction 29
6.3 Experiments on documents about autism 31
6.4 Experimental results of ontology construction in the autism domain 32
6.5 Evaluation of experimental results 36

7 RaJoLink Method for Literature Mining **41**
7.1 The role of rarity in the open discovery process 41
7.2 The role of outliers in the closed discovery process 43
7.3 Method overview 44
7.4 Step *Ra* 45
7.5 Step *Jo* 48

7.6 Step *Link* 49

8 RaJoLink System 53
8.1 The RaJoLink system description 53
8.2 Future development of the RaJoLink system 56

9 Application of the RaJoLink Method to the Literature on Autism 59
9.1 Experimental setup 59
9.2 Experimental results 61
9.2.1 Autism and calcineurin relationship 61
9.2.2 Autism and NF-kappaB relationship 65

10 Evaluation of the RaJoLink Method 71
10.1 Contribution to understanding of autism 71
10.2 Required human effort 74
10.3 Improvements of the RaJoLink's performance based on the evaluation results 79

11 Conclusions 81

12 Acknowledgements 85

13 References 87

Index of Figures 97

Index of Tables 101

Appendix: RaJoLink – User Manual 103

Abstract

Data analysis with machine learning methods, when applied to large collections of text data, enables us to discover new knowledge. This knowledge, once put together, might describe the still unknown connections among phenomena and thus contribute to the formation of new hypotheses in different fields, medicine including. Also, connectivity and computer-supported analysis of numerous large data sets, which include text data, may contribute, in a methodological sense, to the development of e-science. Namely, information that is related across different contexts is difficult to identify with conventional associative approaches. The context-crossing associations, however, are the ones often needed for innovative discoveries. Such associations are called bisociations.

Automated knowledge discovery based on text data sets in the field of medicine is an intriguing challenge as it requires intensive collaboration with domain experts during the processes of both domain-specific text analysis and evaluation. Hence an interactive approach is recommended when text mining and decision support are combined. Also, it is beneficial to apply improved methods of literature mining, searching indirect connections and bisociative knowledge discovery from extensive text databases such as MEDLINE. The major aim here is to unravel the still hidden relations between the researched phenomena and their potential causes. In the process, use of appropriate visualization on the part of the experts is desirable as it supports knowledge discovery and interpretation of results.

The fundamental goal of this thesis is to develop a new methodology for knowledge discovery in text databases that can improve the existing methods of exploring implicit relationships across different domains of expertise by providing a more intuitive computer aided search of unexplored links in literature. To contribute to the current state of the literature-based discovery we designed and implemented an innovative literature mining method for semi-automated discovery of hidden relations that is based on rare pieces of information in a given domain. When these relations are interesting from a medical point of view and can be verified by medical experts, they represent new pieces of knowledge and can contribute to better understanding of diseases.

The developed literature mining method called RaJoLink is intended to support biomedical experts in both open and closed discovery process. In the open knowledge discovery process, hypotheses have to be generated, while in the closed knowledge discovery process, given hypotheses are tested. By identifying relations between biomedical concepts in disjoint sets of articles, the method implements the Swanson's ABC model approach. However, the RaJoLink method analyses such relations in a new way and expands the Swanson's ABC model by suggesting how terms *a* can be determined in advance, as a result of the open knowledge discovery process. The main novelty is a semi-automated suggestion of candidates for agents *a* that might be logically connected with a given phenomenon *c* under investigation. The choice of candidates for *a* is based on rare terms identified in the literature on the topic *c*. As rare terms are not part of the typical range of information, which describe the phenomenon under investigation, such information might be considered as unusual observations about the phenomenon *c*. If literatures on these rare terms have an interesting term in common, this joint term is declared as a candidate for *a*. Linking terms *b* between literature on *a* and literature on *c* are then searched for in the closed discovery process to provide supportive evidence for uncovered connections.

We have applied the RaJoLink method to the scientific literature on autism and have used MEDLINE as a source of data. Autism was selected as the problem domain due to its complexity, insufficient and partial knowledge about its various causes, and because of the strong focus of current medical research towards early diagnosis of this disorder. With the proposed approach we wanted to make a concrete contribution in this direction. In the autism domain we discovered a relation between autism and calcineurin and between autism and transcription factor NF-kappaB, which have been evaluated by a medical expert as relevant for better understanding of autism. To assess the usefulness of RaJoLink in general, we evaluated the potential of our method also in the migraine-magnesium experiment, which represents a gold standard for the literature-based discovery. For all these purposes we also developed a software tool, which implements the RaJoLink method and provides decision support to experts in the process of generating and testing of the scientific hypotheses in biomedical domains.

Povzetek

Analiza podatkov z metodami strojnega učenja omogoča, da iz velikih količin podatkov v podatkovnih bazah izluščimo delčke znanja, ki obravnavani skupaj morda opisujejo še nepoznane povezave med pojavi. Skupaj obravnavana, dotlej nepovezana spoznanja tako prispevajo k novim hipotezam na različnih področjih, med katerimi je že dlje časa tudi medicina. Povezovanje številnih obsežnih tekstovnih virov podatkov ter njihova računalniško podprta analiza prispevajo tudi metodološko k razvoju e-znanosti. Poseben izziv je odkrivanje povezav, ki jih z običajnimi asociacijskimi pristopi ne zajamemo, ker nastopajo v različnih kontekstih. Prav take povezave, imenovane tudi bisociacije, pa so pogosto potrebne za inovativna odkritja.

Odkrivanje znanja iz podatkov na področju medicine zahteva intenzivno sodelovanje z eksperti problemske domene ne le pri vrednotenju rezultatov, temveč tudi že med samo analizo podatkov. Zato je pomemben interaktivni pristop, pri katerem kombiniramo rudarjenje podatkov in podporo odločanju. V sam postopek odkrivanja še neraziskanih povezav med preučevanimi pojavi in možnimi vzroki zanje je smiselno vključiti tudi nove metode rudarjenja besedil. Te omogočajo iskanje posrednih povezav in bisociativno odkrivanje znanja iz izjemno obsežnih tekstovnih baz, kakršna je na primer baza MEDLINE. Za lažjo vključitev medicinskega eksperta je potrebno razviti primerne načine predstavitve vključno z vizualizacijo, kar pospeši izvajanje ciklov odkrivanja znanja in olajšuje interpretacijo rezultatov.

Osrednji namen doktorske disertacije je razvoj nove metodologije odkrivanja znanja iz tekstovnih baz podatkov, ki bo z bolj intuitivnim, računalniško podprtim pristopom izboljšala obstoječe metode raziskovanja implicitnih povezav med pojavi, obravnavanimi v različnih kontekstih. Glavni prispevek k razvoju znanosti na področju odkrivanja znanja iz literature je razvoj in implementacija inovativne metode polavtomatskega rudarjenja po literaturi, imenovane RaJoLink, s katero iščemo dotlej še neodkrite relacije med redkimi izrazi iz besedil v proučevani domeni. V kolikor so taka odkritja zanimiva z medicinskega stališča in lahko eksperti dokažejo njihovo povezavo preko vsebinskih konceptov v literaturi, predstavljajo te dotlej neodkrite povezave vir novega znanja in prispevek k razumevanju obravnavane bolezni.

Metoda tekstovnega rudarjenja RaJoLink je namenjena podpori ekspertom z biomedicinskih področij v njihovem celotnem procesu odkrivanja znanja, tj. pri generiranju in vrednotenju znanstvenih hipotez v raziskovani domeni. Zato vključuje tako zaprt proces odkrivanja znanja, namenjen testiranju hipotez, kakor tudi odprt proces, v katerem hipoteze niso vnaprej poznane. Z odkrivanjem implicitnih povezav med biomedicinskimi koncepti, ki so omenjeni v dotlej nepovezanih člankih, metoda implementira Swansonov ABC model generiranja hipotez, vendar na nov, inovativen način, ne da bi vnaprej poznali ciljni koncept *a*. Ciljni koncept *a* odkrijemo z metodo, kot rezultat samega procesa. Izbira potencialnih kandidatov za ciljni koncept *a* temelji na redkih izrazih, ki jih dobimo v literaturi o problemski domeni *c*. Ker redki izrazi običajno niso tipični za raziskovano domeno, jih lahko obravnavamo kot neobičajne, zanimive informacije o pojavu *c*. Preseke med literaturo o takih redkih izrazih, ki se pojavljajo v strokovnih člankih o preiskovanem pojavu, zato preiskujemo z namenom, da dobimo kandidata za ciljni koncept *a*. Metoda nato v zaprtem procesu odkrivanja znanja išče vezne člene *b* med literaturo o pojavu *a* in literaturo o preiskovanem pojavu *c*, s katerimi bi lahko potrdili novo hipotezo.

V okviru te disertacije smo metodo uporabili na strokovni literaturi o avtizmu, pridobljeni iz baze MEDLINE. Za testno domeno smo izbrali avtizem, ker kljub intenzivnim raziskavam na posameznih področjih še ni dovolj celovitega poznavanja vzrokov te kompleksne motnje, prav tako pa je v medicinskih raziskavah zelo aktualno vprašanje zanesljivega prepoznavanja avtizma že v zgodnjem otroštvu. S predlaganim pristopom želimo konkretno prispevati k temu cilju. Na primeru avtizma smo odkrili povezavo med to motnjo in kalcinevrinom, ki do našega odkritja še ni bila objavljena in je bila medicinsko potrjena kot zanimiv prispevek k razumevanju avtizma. Podobno je bila vzpostavljena tudi povezava s transkripcijskim faktorjem NF-kappaB. Metodo smo ovrednotili še na primeru eksperimenta migrena-magnezij, ki predstavlja klasičen testni primer pri odkrivanju znanja iz literature. Za vse te namene smo razvili programsko orodje, ki implementira metodo RaJoLink in nudi podporo ekspertom pri odločanju v postopku generiranja in testiranja znanstvenih hipotez v biomedicinskih domenah.

Abbreviations

ASD	=	Autism Spectrum Disorders
BDNF	=	brain-derived neurotrophic factor
BoW	=	Bag of Words
GRP	=	gastrin-releasing peptide
MeSH	=	Medical Subject Headings
MMR	=	Measles, Mumps and Rubella
NF-kappaB	=	nuclear factor kappa B
NFAT	=	nuclear factor of activated T cells
TFIDF	=	Term Frequency / Inverse Document Frequency
TNF	=	tumour necrosis factor
XML	=	eXtensible Markup Language

1 Introduction

Internet has become a powerful medium for scientific communication. However, as the amount of scientific articles on the internet has grown so rapidly, the proliferation of electronic scientific publication has become overwhelming. Text mining and knowledge discovery are exciting research areas that play an important role in solving this information overload problem and many researchers have attempted to develop text mining techniques in order to improve the efficiency of knowledge discovery systems. Text mining, which aims at knowledge discovery from text databases (Feldman and Dagan, 1995), is also the principal methodology used in this thesis. In fact, the important aspect of the thesis is the methodological approach that focuses on methods of text mining for knowledge discovery and ontology construction. Besides this, a substantial part of the thesis is devoted to the studies investigating autism, a complex developmental disorder, which is also the application area of our research. Therefore, we focus on using text mining methods for knowledge discovery in biomedical literature.

Scientific progress can be accelerated by knowledge exchange between disciplines to foster new discoveries. Since an abundant quantity of scientific articles is accessible on-line, the usage of large bibliographic databases can support the knowledge discovery process. In biomedicine, databases such as MEDLINE (PubMed, 2008) provide enormous collections of texts that can be used for knowledge discovery. However, finding the right information in broad data collections requires a great deal of skill and time. Consequently, there is a growing need for tools and techniques for processing the vast amount of data available on the internet.

Large advances in text mining and knowledge discovery techniques and tools (Fayyad et al., 1996b; Feldman and Sanger, 2006) facilitate sharing of knowledge and experience among researchers from different, so far not related fields of sciences. This is especially important in interdisciplinary sciences such as biomedicine, since such disciplines need the expertise at the intersections of their component sciences (Shortliffe, 1993). Finding evidence in the biomedical literature to support previously overlooked relations between biomedical concepts can be a way to discover new knowledge. When uncovered relations are interesting from medical point of view and can be verified by medical experts, they can contribute to a better understanding of diseases and related phenomena.

As stated by Arthur Koestler, the bases of scientific discoveries are bisociations (Koestler, 1964). Bisociations, the term coined by Koestler, represent the combination of seemingly unconnected or disparate ideas drawn from distant domains. The idea of performing literature-based discovery of possible relations between previously disjoint concepts was first presented by Swanson (Swanson, 1986; Swanson, 1990). He designed the ABC model to facilitate the discovery of hypotheses by linking findings across scientific literature. The ABC model embodies a search for new indirect relations between two disjoint sets of records (*A* and *C*) via intermediate words and phrases, *B*, that are common to *A* and *C*. According to Swanson, *AB* relations and *BC* relations should have already been separately reported in the published literature, but not considered together. Since our method is based on the same idea, the ABC model approach is described in detail in the related work chapter[1]. Other literature-based discovery methods based on the Swanson's model of discovery are presented as well.

One of the main contributions of this thesis is the development of the new methodology, called RaJoLink, which implements the Swanson's ABC model approach with a new perspective on literature-based discovery. There are traditionally two approaches to literature-based discovery that Weeber and colleagues defined as closed and open discovery (Weeber et al., 2001). In closed discovery, researchers have to start a discovery process by already having generated a hypothesis about the target concept *a*, whereas in an open discovery process the target concepts don't have to be specified in advance. Our literature-based discovery approach combines both, open and closed discovery processes. The main novelty

[1] We use the notations *A*, *B*, and *C* (uppercase symbols) to represent a set of terms (e.g., literature, or set of records, or list of terms), while *a*, *b*, and *c* (lowercase symbols) represent a single term.

of the presented method is a semi-automated suggestion of candidates for agents *a* that might be logically connected with a given phenomenon *c* under investigation. In RaJoLink, the choice of candidates for *a* is based on rare terms identified in the literature on *c*. If literatures on these rare terms have an interesting term in common, this joint term is declared as a candidate for *a*. Linking terms *b* between literature on *a* and literature on *c* are then searched for in the closed discovery approach to provide additional supportive evidence for uncovered connections.

The reasoning underlying the selection of candidates *a* is the following: If there are some rare terms that appear in literature on *c*, let us have a look at all available records about these rare terms. If these records have an interesting joint term *a* in the intersection, let us check if it has some logical connections with *c*. Concentrating on rare terms increases the probability that the suggested candidates have not yet been explored in terms of their connections with *c*.

The need to support the process of communicating research findings across the disciplines has been emphasized also in the context of autism research, which is carried out in different fields, such as behavioural psychology, genetics, biochemistry, brain anatomy and physiology (Belmonte et al., 2004; Zerhouni, 2004). Autism belongs to a group of pervasive developmental disorders that in most cases have an unclear origin. The main characteristic components of abnormal functioning in autism are the early delay and abnormal development of 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* (ASD).

The American National Institutes of Health and National Institute of Mental Health evidenced the lack of studies (Zerhouni, 2004) which would increase the knowledge about risk factors and early development of autism, and that would better define characterization of autism spectrum disorders. The complexity and heterogeneity of autism spectrum disorders, as well as several distinct possible causes pose significant challenges to autism researchers who try to explore causes and to identify phenomena that may lead to autism. These facts have motivated us to select autism as the testing domain for discovering hidden knowledge of value within the scientific literature in different fields.

This thesis contributes a methodology that approaches the above problems. The proposed methodology was introduced in (Petrič et al., 2007; Urbančič et al., 2007) where our study in mining the literature on autism was presented. We continued with the development of RaJoLink as a method in a general form, allowing application in many possible scenarios.

The thesis is structured as follows. Chapter 2 describes the motivation for our research, the hypotheses and the aims of the study and outlines the scientific contributions of this thesis. Chapter 3 presents text mining with related approaches and tasks. This chapter also reviews the work that relates to our research methodology, namely the Swanson's ABC model approach and associated applications, which use the ABC model for discovering complementary concepts in disjoint biomedical articles. In Chapter 4 we provide the necessary background knowledge about the autism domain. Chapter 5 presents the experimental design, the materials and concepts used for the present thesis. Chapter 6 gives an overview on ontologies and presents our studies of structuring domain knowledge by construction of ontologies. In Chapter 7 we explain the RaJoLink method for identifying implicit and previously unknown connections on the basis of rare terms. Chapter 8 presents the RaJoLink system that applies the RaJoLink methodology and gives directions for further development of the RaJoLink system. Chapter 9 illustrates the application of the RaJoLink method to the literature on autism. We describe our literature mining results by example pairs of implicit connections that we managed to identify from biomedical articles. Experimental results given in this chapter are followed by their evaluation in Chapter 10 that includes also the method’s evaluation on the migraine-magnesium example. In the final chapter we summarize the overall conclusions of this work.

2 Aims and Hypothesis

The primary focus of this thesis is on the development and application of a methodology and a software tool for text data analysis with the objective of facilitating the process of knowledge discovery from text documents. This chapter highlights important background issues of this research applied to the medical domain of autism. In particular, our motivation, propositions and objectives behind this work are given. In order to exploit existing but often unconscious and overlooked knowledge that is hidden in scientific articles we investigated the potential of two information science disciplines: the text mining and link analysis methods. As a result, we aim at specifying the contribution of this thesis, both to the field of information science and to the understanding of autism.

2.1 Motivation

Scientific research is increasingly relying on the internet as an information source and communication medium due to the fast information technology development and in particular due to the rapid growth of the internet access and its widespread use. Scientists have to cope with information systems and heterogeneous data sources that are constantly increasing. Frequently, they have to go through a large number of records retrieved from huge databases before they discover a relevant piece of knowledge.

We explore the possibilities to improve the current knowledge discovery approaches using the text mining techniques. As a series of text mining researches have already been conducted, we carry out the comparison of the existing techniques to identify the issues that have the greatest potential for scientific progress in the literature-based discovery area.

In accordance with the motive of focusing on the literature-based link discovery, the original idea of the dissertation was to investigate how to use the text mining tools to support the analysis work of medical researchers in the autism domain and other scientists as well. Scientific medical knowledge that is embodied in text such as journal articles and conference proceedings has become widely accessible to the scientific community, as well as to the general public due to the growing popularity of the internet. Knowledge discovery from text data is an increasingly important area of research driven by the internet growth and public access to very large digital libraries. The work described in this thesis is part of this new research stream on literature-based knowledge discovery and focuses on particular area of link discovery by mining public knowledge that has not been explored before.

As an important component of the knowledge base in our research methodology we use ontologies constructed from the application domain literature. Such domain ontologies contain background information and define the concepts and the structure of a conceptualization of a particular target domain. We are confident that ontologies help domain experts in their cognitive processes to properly model the conceptual understanding of the domain under investigation. Therefore, we use ontologies to provide an integrated view over knowledge fragments that otherwise would remain known only to individuals within a particular research community. When the knowledge engineer captures the knowledge of a domain into ontology, we propose ontologies to be used in collaboration with domain experts who can help identifying some specific knowledge concepts that fall within the scope indicated by these experts. Moreover, ontologies also facilitate communication between the knowledge engineer and domain experts and make the understanding and interpretation of the domain literature easier.

The idea for our research in autism domain emerged in the early spring of 2006 when Slovene parents of children suffering from autism launched a collective campaign to sensitise the wider society on the problems that children with autism and their families face on a daily basis. Through this campaign a non-governmental organisation called Centre for Autism urged the Children's Rights Department of the Slovene Ombudsman's Office to speak about the scarce resources and knowledge about autism in order to raise the public awareness of autism in Slovenia and in other European countries (European Network of Ombudspersons for Children, 2006).

At the end of 2007 the United Nations General Assembly decided to designate the 2nd of April as World Autism Awareness Day, to be observed every year beginning in 2008. Therefore all member states, relevant organizations of the United Nations system together with other international organizations, as well

as civil society, including non-governmental organizations and the private sector were invited to observe World Autism Awareness Day in an appropriate manner, for the purpose of improving health care, education, training and intervention for children with autism (United Nations, 2007). The growing international concern regarding autism motivated our research, which seeks to understand the autism issues and to enable the design of a methodology that can improve the knowledge discovery in complex domains, such as autism.

We started our investigation of autism by mining the published literature on autism and by consulting medical doctor Marta Macedoni-Lukšič, a developmental paediatrician who is also a foundress and directress of the Institute for Autism and Related Disorders in Ljubljana. After first publications of our research results in October 2006 (Petrič et al., 2006a; Petrič et al., 2006b) we decided to stick to the research on autism with the aim to improve our understanding of this disorder and to further explore this area of interest. The present thesis therefore aims to gain new knowledge and more insight into the autism domain by literature analysis.

This thesis suggests a new approach to the literature-based knowledge discovery by:

- Developing a novel knowledge discovery methodology for processing and analysing text data that will improve the existing methods of literature-based link discovery by providing a more intuitive search of unexplored links between information fragments;

- Placing domain ontologies in the framework of the literature-based knowledge discovery to facilitate the acquisition of insights and understanding of a research domain and especially for the better communication between the domain expert and the knowledge engineer;

- Yielding a better understanding of the various aspects of autism by semi-automatic construction of autism ontologies based on the scientific literature;

- Casting more light on the phenomena of autism by discovering and exploring some novel factors that may be helpful in elucidating the nature of autism.

2.2 Research hypotheses

Text mining has been shown to be useful for discovering of new, previously unknown information, by automatically extracting information from different text resources (Feldman and Sanger, 2006). However, from a knowledge discovery science perspective, there is still a lack of research studies investigating the literature-based approach towards link discovery and the specifics of such practices. It would thus be interesting to make a systematic examination of alternative techniques for discovering links and relations in literature through text mining.

Besides this, there is a need to identify critical pathways on a hypotheses generation level of the knowledge discovery process in order to contribute to the selective formation of plausible scientific hypotheses in a domain of interest. The presented realizations give rise to the following research hypotheses that underlie this dissertation:

- The basic hypothesis is that bisociations, representing novel interesting connections between distant rarely mentioned research findings, can be extracted from the published biomedical literature. We suppose that they can add relevant new knowledge to complex biomedical domains such as autism through the intermediate links in the chain of events. In this context we suggest that what is common to the rare terms from a research domain literature indicates a subject that is implicitly connected with the domain of discourse.

- One further hypothesis of this thesis is that given two implicitly linked subjects by an assumption, this assumption can be verified by neighbouring documents in the documents' similarity graphs. Our argumentation is that outlying documents of two implicitly linked subjects can be used to search for relevant linking terms between the two subjects.

- Consequently, we assume that a novel approach to literature-based link discovery, which is

based on the principle of rare terms together with the notion of bisociation is going to improve the support to user-guided knowledge discovery. This is expected as a result of a different point of view in the text mining approaches with the proper cross-disciplinary consideration of the research domains in order to overcome the barriers of traditional thinking paradigms that are dominant within disciplines.

Moreover, since literature analysis can be very time-consuming, we believe that the literature-based research to the generation of scientific hypothesis can be speeded up by an automated system based on innovative text mining methods.

2.3 Scientific contributions

The scientific value of the present thesis is twofold. On the one hand, we made significant contributions to the field of knowledge discovery from text documents and at the same time we contributed to the understanding of autism. In particular, the contributions of this thesis are as follows:

- We improved the open literature-based discovery process with an interdisciplinary, semi-automated approach to hypotheses generation that bridges the overspecialization in sciences. The current literature-based discovery systems depend fundamentally on the word co-occurrences in text. Instead, we proposed a radical shift in thinking and decided to search for associative concepts by mining the words that rarely appear in the documents about the domain under examination. Rare terms are thus our innovative way in the open discovery process towards the candidates for term *a*.

- We made important progress also in the closed discovery process by focusing on outlying and their neighbouring documents in the documents' similarity graphs. We demonstrated that outlying documents could be used as a heuristic guidance to speed-up the search for the linking terms and alleviate the burden on the expert when hypotheses have to be tested. We showed that with the similarity graphs that enable the visual analysis of the literature it is easier to detect the documents, which are very interesting for a particular link analysis investigation for the reason that such outlying documents often represent particularities in domain literature.

- We observed a particular category of link structures, namely the context-crossing associations, called bisociations (Koestler, 1964). We employed them in combination with the so called intermediate links (Swanson, 1988) that serve to bridge the disjoint and non-interactive domains. Bisociations and intermediate links are undoubtedly extremely important components of the scientific research and we proved that combining them together enhances the literature-based knowledge discovery.

- We developed a method supporting the overall process of open and closed knowledge discovery, which we called RaJoLink after its fundamental procedural elements: *Ra*re terms, *Jo*int terms, and *Link*ing terms. To address the need for fast text retrieval and text analysis in practice, we constructed a software tool RaJoLink which implements the RaJoLink method and helps biomedical experts to automatically investigate yet undiscovered relations between concepts from disjoint research subfields.

- We examined the use of a particular terminology within the literature on autism and emphasized the areas in which the autism research community has been involved. Our ontologies of the autism domain reflect the general knowledge and understanding of autism as it is documented in scientific articles. In this manner, our survey of the autism research represents a worthwhile contribution to community-wide efforts to acquire knowledge about autism. This way we provided also a general framework that enables the biomedical researchers in the fast-growing fields of science to keep abreast of current developments.

- We managed to identify the novel candidate targets to elucidate the neurobiological mechanisms that underlie the autistic spectrum disorders. As a matter of fact, our study calls

attention on two indices, both of which we associated with conditions seen in patients with autism. The first one is calcineurin, the major Ca2+-dependent phosphatase in neurons. The second one is nuclear factor kappaB, a transcription factor that plays a pivotal role in immune and inflammatory responses.

The early papers on this research were accepted in 2006 for the 9th International Multiconference Information Society, IS 2006 (Petrič et al., 2006a; Petrič et al., 2006b) and in 2007 for the 11th Conference on Artificial Intelligence in Medicine, AIME 2007 (Urbančič et al., 2007). Our findings were published also in the Journal of Biomedical Informatics (Petrič et al., 2009), which presents high-quality original research papers and reviews by world-renowned scientists in the area of biomedical informatics. The following papers show the development and highlight the main scientific contributions of this thesis to the research in text mining and autism:

- Petrič, I.; Urbančič, T.; Cestnik, B. Comparison of ontologies built on titles, abstracts and entire texts of articles. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O.; Musek, J.; Osredkar, M. J.; Kononenko, I.; Novak Škarja, B. (eds) IS-2006. Proceedings of the 9th International multi-conference Information Society. pp 227-230 (Ljubljana, Slovenia, 2006).

- Petrič, I.; Urbančič, T.; Cestnik, B. Literature mining: potential for gaining hidden knowledge from biomedical articles. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O.; Musek, J.; Osredkar, M. J.; Kononenko, I.; Novak Škarja, B. (eds) IS-2006. Proceedings of the 9th International multi-conference Information Society. pp 52-55 (Ljubljana, Slovenia, 2006).

- Petrič, I.; Urbančič, T.; Cestnik, B. Discovering hidden knowledge from biomedical literature. Informatica 31(1), pp 15-20 (2007).

- Urbančič, T.; Petrič, I.; Cestnik, B.; Macedoni-Lukšič, M. Literature mining: towards better understanding of autism. In: Bellazzi, R.; Abu-Hanna, A.; Hunter, J. (eds) AIME 2007. Proceedings of the 11th Conference on Artificial Intelligence in Medicine in Europe. pp 217-226 (Amsterdam, The Netherlands, 2007).

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

- Petrič, I.; Urbančič, T.; Cestnik, B.; Macedoni-Lukšič, M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. Journal of Biomedical Informatics 42(2), pp 219-227 (2009).

The six papers at the core of this thesis describe our research on three distinct but related topics in information science. One of them is the computer-supported biomedical literature research, which is a topic of intense interest among information scientists, since the biomedical literature is an extremely rich information source but on the other hand identified as information overload experienced by the biomedical scientists. The particular problem domain in this regard is that of autism that we chose for the application of our methodology. A further topic of interest is the construction of ontologies that can provide the necessary background knowledge and serve as a means towards link discovery in the closed discovery process. Another research topic, which is of particular importance for this thesis and thus the main centre of action, is text mining for knowledge discovery in the open discovery process. The following survey of related work therefore contains an overview of the disciplines of text mining and link analysis with specific reference to their application to the biomedical domains.

3 Related Work

The intention of this chapter is to survey previous research in text mining and link analysis in order to put our work into its proper context. In particular, we look at advantages and disadvantages of approaches that concerned knowledge discovery in existing biomedical bibliographic databases and are based on the Swanson's ABC model. Therefore, we review selected publications related to the topics covered in this thesis and discuss the main issues focusing on methodological aspects of text mining and link analysis.

3.1 Text mining and knowledge discovery

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.

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. The situation becomes even more striking when a person wants to obtain an insight into a field that does not fall directly into his or her area of expertise. 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., 1996b). 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. Likewise, when a set of articles serves as a source of data, the process is typically called literature mining. In Figure 1 we illustrate the main phases of the text mining process, where the Bag of Words approach (Sebastiani, 2002) is used for representation of collection of words from text documents disregarding grammar and word order.

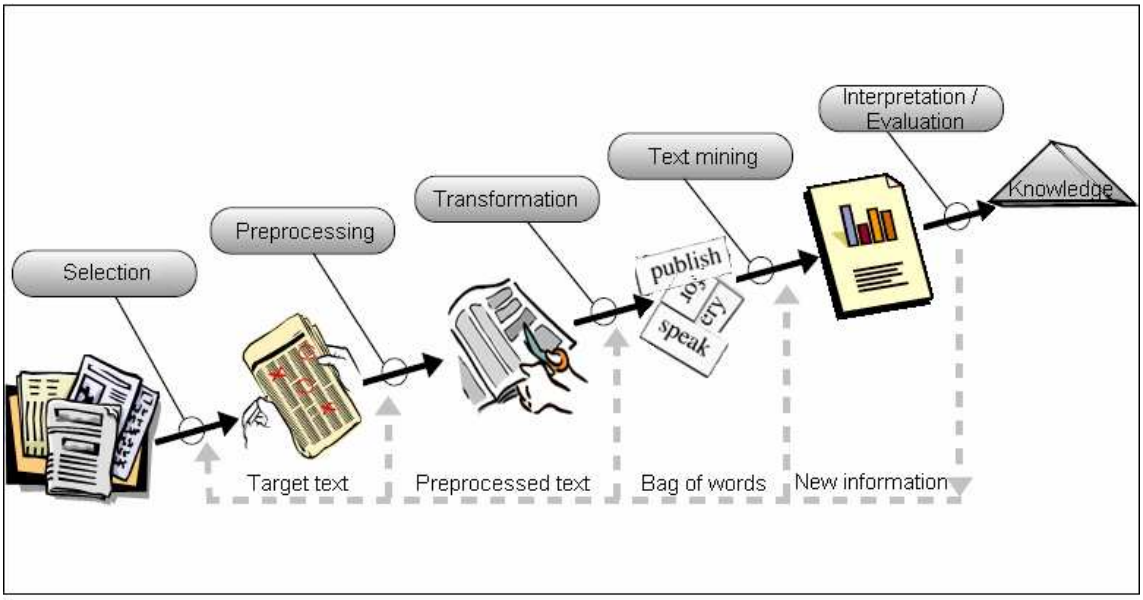


Figure 1: *Text mining process*. The sequence of steps is modelled in conformity with a definition of knowledge discovery in databases (KDD) process as originally proposed by Fayyad and colleagues.

We modelled the sequence of the text mining steps according to the basic flow of knowledge discovery steps as defined by Fayyad, Piatetsky-Shapiro and Smyth (Fayyad et al., 1996a). Text mining applied to texts of a biomedical domain is named also biomedical text mining. The purposes of such text mining are to efficiently identify needed information, to uncover relationships hidden by the large amount of available information, and for the most part to provide researchers with automated computer methods that can simplify the information overload (Cohen and Hersh, 2005).

Although the technology for data and text mining is well advanced, its potential still seems to lack sufficient recognition in practice. 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 and 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 and 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 emphasize the need for increasing the efficiency of communication of research findings to the related science community. That is also a reason why we have examined the text mining potential on the literature on autism.

The amount and the growth speed of scientific information that is available online have strongly influenced the way of work in the research community which calls for new methods and tools to support it. Biomedical field is a very good example, with MEDLINE database, the primary component of PubMed (the United States National Library of Medicine's bibliographic database), which covers approximately 5,200 journals published in more than 80 countries, contains more than 16 million citations from 1949 to the present (Figure 2), and increases for more than 2,000 complete references daily (Pubmed, 2008). Figure 2 shows the numbers of documents cited in MEDLINE each year since 1948 according to the MEDLINE citation counts by year of publication (U.S. National Library of Medicine, 2008).

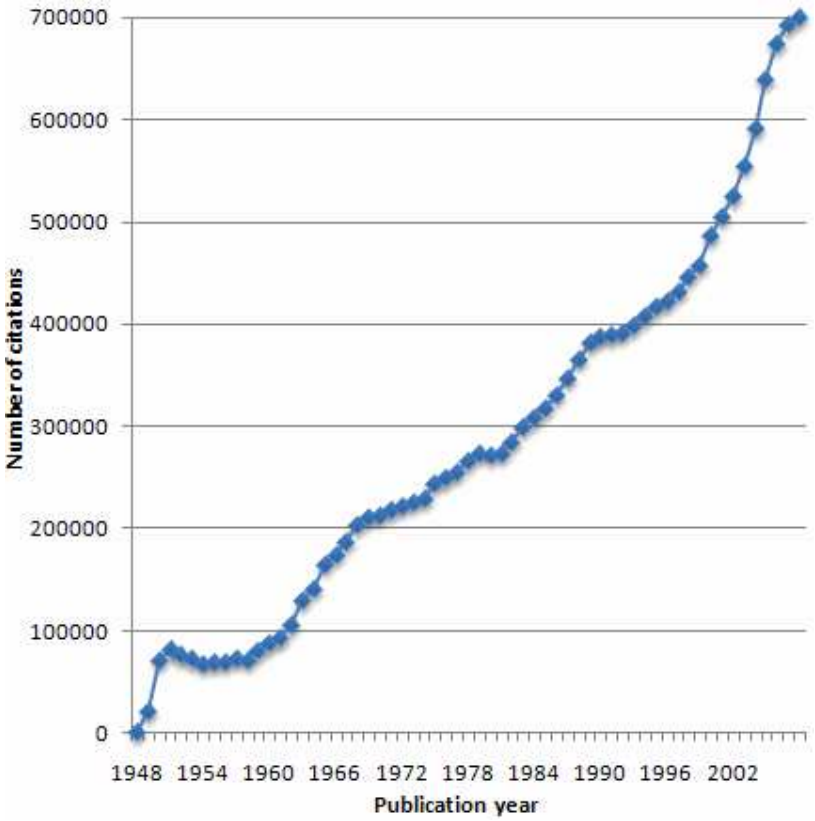


Figure 2: *Citation counts of MEDLINE publications*. The numbers of documents cited in MEDLINE in the period from 1948 to 2007 (Source: U.S. National Library of Medicine, April 2008).

Knowledge technologies and among them especially knowledge discovery based on data mining and text mining, offer new possibilities by their ability to uncover hidden relationships in data (Fayyad et al., 1996b). There are several examples of European research projects, where data mining has been successfully applied in biomedical domains. Various data mining research activities are integrated into the Information Society Technologies Work Programme and are in this way part of the so-called Framework Programmes of the European Union.

European projects funded under the Sixth Framework Programme for the years 2002-2006 involved the following data mining research activities for biomedical applications (European Communities, 2008):

- Data mining and statistical methods to investigate the content of the scientific literature and exploit the synergy between bioinformatics and medical informatics;
- Context-based information integration achieved by text mining and link discovery applied to the evidence-based medicine, literature and patent mining, and molecular biology, with the purpose to study infectious diseases;
- Data mining and analysis of the biomedical data from viruses, patients and literature resulting in a rule-based decision support system for drug ranking and knowledge discovery in medicine;
- Semantic modelling, fuzzy inference and data mining of clinical and genetic data for translation of medical concepts into syntactic values that facilitate the study of cervical cancer;
- Information fusion and data mining of biomedical data for knowledge discovery, early diagnosis, screening, disease prevention, treatment and follow up of children patients with heart diseases, inflammatory diseases and brain tumours;
- Semantic interoperability and data mining in biomedicine including the fields of genomics and proteomics to facilitate knowledge transfer between different scientific disciplines and to support cooperation between the academic community and organizations in the health sector;
- Collaborative data mining, modelling, and visualization of clinical, demographic, and patient-specific genomic and proteomic data in order to link heterogeneous data about metabolic diseases and cardiovascular risks;
- Data mining of heterogeneous clinical data as a part of an agent-based distributed decision support system for the diagnosis and treatment of brain tumours;
- Clinical data mining as a support in developing a simulation model of the MAP-kinase pathway in relaying signals from the plasma membrane into the nucleus, which is required for rational anti-cancer therapies.
- Genomic and proteomic data mining for discovery of mutations on tumour suppressor genes that contribute to the development of colon cancer;
- Data mining grid services and ontology based semantic integration of clinical, genomic and proteomic data in order to support complex knowledge discovery in the cancer research area.

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. The principal feature of text mining is its concentration on the document collection, which can be any group of text-based documents (Feldman and Sanger, 2006). Essentially, text mining is used to denote the analyses of large quantities of natural language text and the detection of usage patterns with the goal to extract probably useful information (Sebastiani, 2002). 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 and 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 (Segura-Bedmar et al., 2008),
  - gene and protein names within a collection of MEDLINE abstracts (Tanabe and Wilbur, 2002).

  Text mining approaches: lexicon-based, rules-based, statistically based, combined.

- *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 and other biological names synonyms (Hirschman et al., 2002),
  - biomedical term abbreviations (Schwartz and Hearst, 2003).

  Text mining approaches: combination of named entity recognition with statistical, support vector machine classifier-based, and automatic or manual pattern-based matching rules algorithms.

- *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 (Chen et al., 2006),
  - texts containing a certain type of information (Liu et al., 2004).

  Text mining approaches: classification rule induction, computation of distances between keywords, and support vector machine classifier-based.

- *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 (Giles and Wren, 2008),
  - text-based gene clustering (Raychaudhuri et al., 2002).

  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 by multiple queries (Becker et al., 2003),
  - biomedical terminology recognition and clustering in an integrated framework (Nenadic et al., 2003).

  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 (Srinivasan and Libbus, 2004),
  - potential new uses and therapeutic effects of drugs (Weeber et al., 2003).

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

3.2 Link analysis

While it is often easy to automatically collect data, it requires considerable effort to link and transform them into practical information that can be used for discovering new knowledge. The process of ideas and knowledge diffusion from one society to another can be carried through a particular terminology used for specific subject matters and through citations channels. Link analysis can reveal the structures of implicit communities and suggest interesting points in such communications channels for knowledge sharing. It enables building up networks of interconnected objects by exploring associations among objects and linking them through the relevant relationships.

Link analysis has developed over the past twenty years in many different subjects, from computer science and informatics to social sciences. The most characteristic for the link analysis field is the category of graph representations of objects and links between them such as those illustrated in Figure 3. In fact, graph data structures can be widely applied to social research issues, such as social network analysis. Therefore, they are often employed to investigate linking patterns in various social networks, including mailing lists and online social networks of communities of people or of business with common interests.

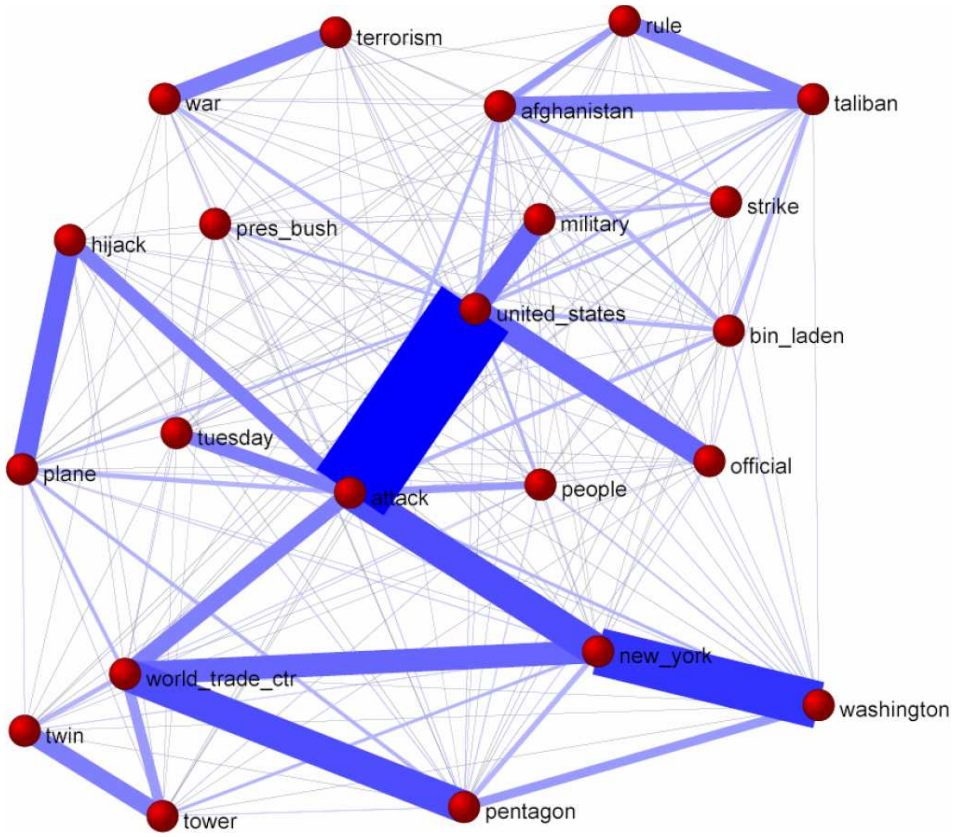


Figure 3: *Terror news network analysed by Pajek (a popular link analysis tool) based on news reports following the September 11 attack on the United States (Batagelj and Mrvar, 2003)*. In the example network, the thickness of an edge represents the frequency of co-appearance of two words that are linked by the given edge (Source: http://vlado.fmf.uni-lj.si/pub/networks/pajek/pics/examples.htm, January, 2008).

Link analysis techniques have frequently been applied to issues such as co-citation analysis to explore relationships between authors or as co-coupling measurements to determine the similarity between documents. McCain, for example, used the author co-citation analysis to display dynamics in intellectual structure of macroeconomics by focusing on changing relationships among authors in this field of economics over time (McCain, 1984). A number of link-based algorithms have been proposed also for the web link analysis to identify web spam (Becchetti et al., 2008).

In computer science, the link analysis functionality was incorporated in several web information retrieval algorithms e.g., the Hypertext Induced Topic Search (HITS) algorithm that was developed by

Kleinberg (1999). The HITS algorithm (Kleinberg, 1999) is a link analysis algorithm that views the web as a graph where web pages are nodes. It ranks web pages by utilizing the hyperlink structure of the web. The ranking is based on the authority and hub value. The hub value estimates the quality of outgoing links from the page to other pages. Therefore, a good hub is a page that refers to many good authority pages. The authority value is the sum of the hub values of all incoming links (i.e. of the web pages that point to the page) where a good authority is a page that is pointed to by many good hubs. The authority value is thus used to estimate the value of the content of a web page.

Similarly, in information science, there is a growing interest in link analysis for studying the structure of hyperlink networks of documents (e.g., web pages), categories, and users of common interests (Thelwall, 2004). Such link analysis studies typically apply data and text mining algorithms to large collections of web data. Regarding web research, there is a particular link structure analysis called webometrics (Almind and Ingwersen, 1997). Webometrics has emerged as a research field of information science in recent years. It examines the quantitative aspects of how different users access and handle information in different contexts.

Some programs for large network analysis are in addition powerful visualization tools. A notable representative of such tools is Pajek (Figure 3) that was designed by Batagelj and Mrvar to support link analysis in different areas: internet networks, citation networks, diffusion networks, biomedical and genomics network structures (e.g., organic molecules or protein-receptor interaction networks), genealogical networks (e.g., marriages or lines of descendants) and many others (Batagelj and Mrvar, 1998).

In the example shown in Figure 3, Batagelj and Mrvar aimed to analyse the presence of links between terrorists on the basis of news reports following the September 11 attack on the United States that were published during 66 consecutive days (Batagelj and Mrvar, 2003). With this purpose, they obtained the Reuters terror news network from the CRA networks, which was produced by Corman and Dooley at Arizona State University and subsequently transformed into the Pajek format by Batagelj (Batagelj and Mrvar, 2003). Networks like the example illustrated in Figure 3 are composed of set of vertices, which represent social entities and the set of lines between them, representing the relationships in the network. The vertices of this terror news network are words. If two words appear in the same text unit (i.e., in a sentence) then an edge is drawn between them.

The popular link analysis tools, such as Pajek, which are designed for building up networks of interconnected objects, normally need to operate with structured data. Therefore, datasets have to be prepared in particular machine-readable formats. On the other hand, more specific text mining techniques have to be used for the fully automated link analysis of unstructured data such as text documents from large bibliographic databases. In the continuation, we concentrate on mining of such literature to generate hypotheses as a central point of our research interest. Accordingly, we present the literature mining by link analysis as a particular text mining approach towards integration of real problem analysis and extraction of potentially useful information from literature.

3.3 Creativity through bisociations

One of the main challenges in literature mining is the extraction of implicit and previously unknown interesting information from scientific articles. Many scientific discoveries involve the hyptothetico-deductive pattern of thinking (Lawson 2002). Lawson describes the basic elements of hyptothetico-deductive science, as can be seen, for example, in Galileo's discovery of Jupiter's moons. The elements involved are: making a puzzling observation, identifying a causal question, formulating hypotheses and using them to generate expected results, making actual observations and comparing them with the expected ones, and finally drawing conclusions. Analysis of implicit associations hidden in the scientific literature can guide the hypotheses formulation and lead towards discovery of new knowledge.

The theoretical framework of our literature-based hypotheses generation research is grounded in the associationist creativity theory (Mednick, 1962) with particular focus on the unanticipated context-crossing associations that Koestler called bisociations (Koestler, 1964). According to his theory, a bisociation is the link between two concepts from commonly unrelated domains that joins such domains by observing them from a particular, innovative point of view. Bisociations have the potential of generating radical discoveries, by enabling the entirely new cross-disciplinary connections among concepts from those contexts that are normally considered as distinct categories. Besides scientific discovery, the bisociations can enhance also the humour creation techniques and inspire the artists to create original works of music, theatre and art (Koestler, 1964).

Mednick (Mednick, 1962) defined creative thinking as the ability to generate new combinations of distant associative elements (e.g., words). He explained that thinking of concepts, which are not strictly related to the elements under research, inspires unexpected useful connections between elements and thus considerably improves a creative process. Actually, marginal observations are not necessarily characterized by mistakes or inaccuracies but may provide an indication of valuable information (Barnett and Lewis, 1994). From this point of view, creative thinking constantly involves a process of evoking latent possibilities to discover new useful information and unforeseen knowledge.

In the light of bisociative thinking, the creation of new discoveries lies in linking the previously independent concepts from two traditionally disparate frames of reference (Koestler, 1964). Therefore, the more independent are the concepts under observation, the more outstanding will be the novel discovery of bisociation. By such definition, the logic of bisociations differs from the classical associative thinking that according to Koestler refers to previously established connections among ideas, which have been however hidden until the new discovery. In contrast to bisociations, where no comprehensive computer aided methodology has yet been developed on this basis, the associative contexts have been analyzed by many researchers also in the literature mining area. The literature-based knowledge discovery process as proposed by Swanson (Swanson, 1986) was extensively studied.

3.4 Swanson's model of knowledge discovery

The bibliographic databases such as MEDLINE can serve as a rich source of hidden relations between biomedical concepts, as shown in 1986 by Swanson (Swanson, 1986). Swanson regards scientific articles as clusters of somewhat independent sets of literatures, where common matters are considered within each set (Swanson, 1990). He calls *noninteractive* those literatures that do not cite one another, that have no articles in common, and that are not cited at the same time by other papers. According to his proposal, such distinct unrelated literatures could be linked to each other by arguments that they treat. Consequently, if two literatures can be logically related by arguments that each of them addresses the unobserved connections between them represent potential sources of new knowledge. For instance, if literature *A* (i.e. a set of all available records about *a* in the database serving as a source of data) reports about term *a* being in association with term *b*, and another literature *C* associates term *c* with term *b*, we can thus assume literature *B* to be an unintended implicit potential connection between literatures *A* and *C*.

When concentrating his attention on finding implicit relationships between literatures *A* and *C* Swanson found out that terms *b* can be employed in at least two different manners. The first is to discover novel *AC* relationships, the second is to use existing terms *b* and discover a novel combination of their joining properties (Swanson et al., 2006).

By considering unconnected sets of articles Swanson managed to make several surprising discoveries. In one of them, he discovered a relationship between Raynaud's syndrome and dietary fish oil (Swanson, 1986). Until this discovery, the literature about Raynaud's syndrome and the literature addressing fish oil were disjoint as they had no common authors or mutual citations. Swanson discovered connections between these two literatures by reading biomedical literature for two distinct reasons. On the one hand, he was interested in clinical tests that evidenced positive effects of dietary fish oil on blood, arteries and heart such as, for example, inhibited platelet aggregation, reduced blood viscosity and decreased low-density blood lipids. On the other hand, he noticed many articles on Raynaud's disease reporting on unusually increased blood viscosity and augmented platelet aggregation. Thus he could hypothesise that patients with Raynaud's syndrome might benefit from consuming dietary fish oil.

Similarly, while studying two separate literatures, the literature on migraine headache and the articles on magnesium, he found implicit connections that were unnoticed at the outset of his research (Swanson, 1990). Swanson noticed the possible relationship between the disjoint literatures on migraines and on magnesium by the intermediate literature. In fact, some linking terms, such as *calcium channel blockers* and *spreading cortical depression* appeared frequently in the titles of both the migraine literature and the magnesium literature. However, prior to the Swanson's discovery, a few researchers (e.g., Altura, 1985) had given attention to a direct magnesium-migraine connection, but laboratory and clinical investigations started numerously only after the publication of the Swanson's convincing evidence.

Swanson investigated the process of finding implicit connections between disjoint literatures using the titles of articles and their Medical Subject Headings (MeSH) terminology, which provides descriptors for MEDLINE records in a hierarchical structure (Nelson et al., 2001). For literature-based discovery, Smalheiser and Swanson (Smalheiser and Swanson, 1998) designed the ARROWSMITH system based on MEDLINE search. Their major focus was on the hypothesis testing approach (Swanson et al., 2006), which

Weeber and colleagues (Weeber et al., 2001) defined as a *closed* discovery process (left model in Figure 4), where both *a* and *c* have to be specified at the start of the process.

As reported by Weeber (Weeber, 2007), Swanson's first literature-based hypothesis that dietary fish oil might benefit patients with Raynaud's disease (Swanson, 1986), was a coincidence. Thereafter, Swanson studied the literatures of both target concepts, namely fish oil (*a*) and Raynaud's disease (*c*) for finding the linking terms (*b*) with already having this hypothesis in mind. On the other hand, an *open* discovery process (the right model in Figure 4), is characterized by the absence of requirement for advance specification of target concepts. If we are investigating a subject denoted with term *c*, the open discovery starts with having only term *c* and the corresponding set of articles in which term *c* appears (called also literature *C*), without knowing target term *a*, which is discovered later as a result of this process.

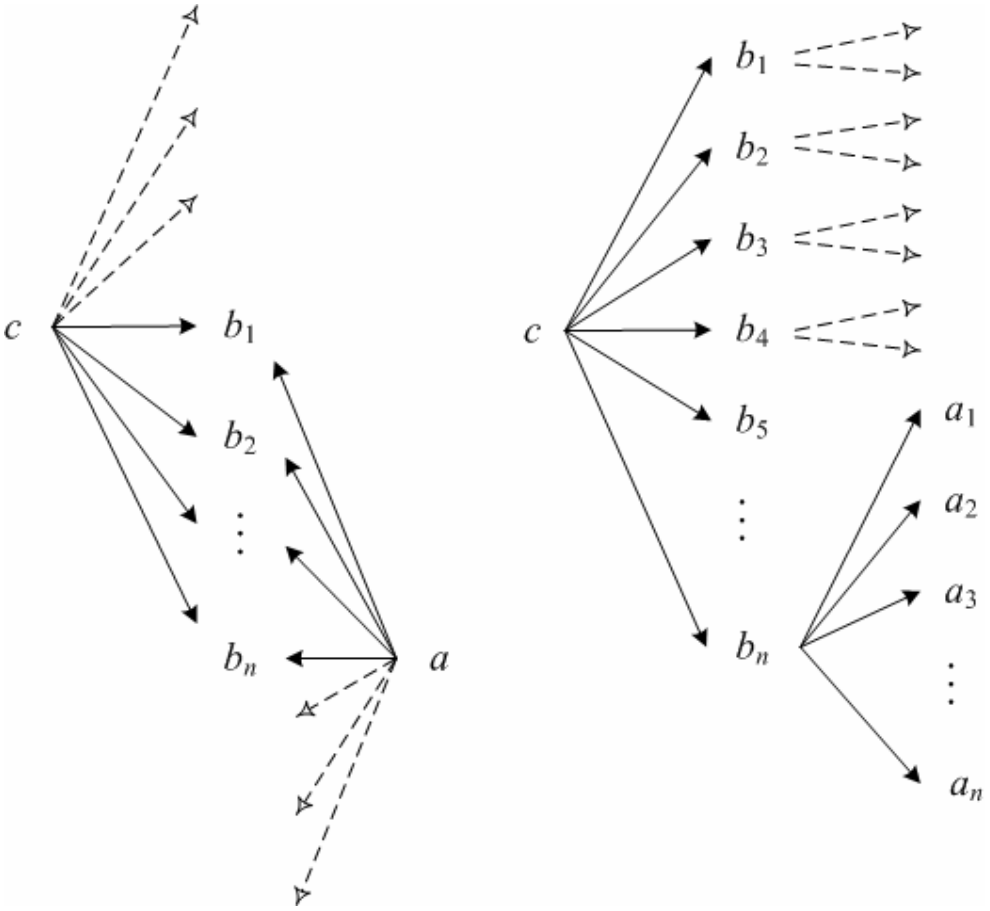


Figure 4: *Closed (left model) versus open (right model) discovery process as defined by Weeber et al. (Weeber et al., 2001).*

3.5 Approaches based on Swanson's ABC model

Several other text mining and hypothesis generating systems supporting literature-based discovery were developed following the early work of Swanson. Lindsay and Gordon (Lindsay and Gordon, 1999) tested Swanson's discoveries of connecting Raynaud's disease to fish oil and migraine to magnesium deficiency by using different lexical statistics, such as word frequency counts. The authors tried to find relevant words on top of ranked lists in their open discovery approach and thus replicated Swanson's first two discoveries. However, their relative frequency statistic failed in suggesting magnesium and extensive analysis had to be based on human knowledge and judgement rather than automated procedures.

Weeber (Weeber, 2007) pointed out that the expert knowledge is indispensable in the literature-based discovery to choose among possible results and to determine potentially contradicting information. Weeber and colleagues (Weeber et al., 2001) simulated the same two Swanson's discoveries with Natural Language Processing techniques by searching biomedical Unified Medical Language System (UMLS) concepts (U.S. National Library of Medicine, 2006b) in texts. They developed a system for generating new hypotheses

from the literature, called Literaby (Weeber, 2007). The system identifies the concepts in the UMLS that are related to a starting term *c* and executes a MEDLINE query. Consequently, linking terms *b* are extracted from titles and abstracts of the resulting citations from MEDLINE and selected by the user who should have the expert knowledge in the domain of interest. At this stage, the filter with 134 semantic categories can help the user to choose the most promising *b* terms from the list of potential linking terms. Then again, a MEDLINE query is performed with the selected *b* terms to find their co-occurring concepts (*a* terms). Once more, the user forms a semantic filter by setting among the 134 semantic categories to retrieve the most promising target terms. The generated hypotheses are then evaluated in the closed discovery phase, where both literatures on *A* and *C* are downloaded form MEDLINE and analyzed to find interesting linking *b* terms. The most plausible are those associations between literatures *A* and *C* that have the highest number of overlapping *b* terms.

The Srinivasan and colleagues' (Srinivasan et al., 2004) discovery approach, on the other hand, relies nearly completely on Medical Subject Headings (MeSH). Their open discovery approach is established on topic profiles (Figure 5), where topics represent subjects of interest that are derived from the inspected text collection. As the text collection is obtained from MEDLINE, the topic profiles are vectors of weighted MeSH terms. Term weights calculated for the MeSH terms are a modification of the standard TF*IDF (term frequency inverse document frequency) scores. As MeSH terms are classified with one or more of the 134 UMLS semantic types, this advantage of MeSH metadata is taken to restrict the search process and to consider only MeSH terms belonging to certain semantic types. The approach begins by building the topic profile for the starting literature *C* and restricts the MEDLINE search to only those MeSH terms that belong to certain semantic types specified by the user. For each semantic type the top ranked MeSH terms are automatically selected, which represent the *b* terms. After that, profiles are built for each of the *b* terms and analysed in combination to select a candidate target term *a*. Although Srinivasan and colleagues reduced the amount of manual effort and intervention, an important part of the open discovery process depends almost entirely on the user. For example, the user needs to choose the *AC* pairs of interest and has to research the related literatures *A* and *C* in order to find the supporting evidence.

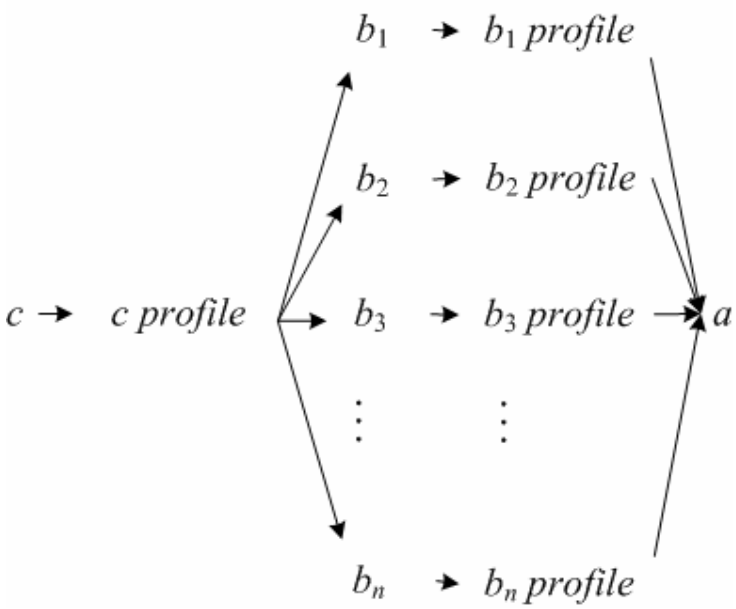


Figure 5: *Open discovery process as applied by Srinivasan and colleagues.*

Some other researchers concentrated on literature mining and knowledge extraction from biomedical databases, too. Among them, Hristovski and Peterlin constructed a literature-based biomedical discovery support system, called BITOLA, and found the evidence for associations between several genes and diseases (Hristovski et al., 2005). BITOLA applies association rule mining to find novel relations between literatures. Another literature-based discovery system, named LitLinker (Figure 6), was presented by Yetisgen-Yildiz and Pratt (Yetisgen-Yildiz and Pratt, 2006), who combined knowledge-based methodologies with statistical methods to capture new connections between diseases and chemicals, genes or molecular sequences from biomedical literature. To recognize the linking and their correlated terms, LitLinker follows a statistical approach supported by the background distribution of term probabilities. First, it finds the linking terms (*b* terms) that are directly correlated with a starting term *c*. Next, it searches

for target terms (*a* terms), which are associated with each linking term. Finally, it ranks the target terms by the number of linking terms *b* that correlate the resulting target term *a* with the starting term *c*.

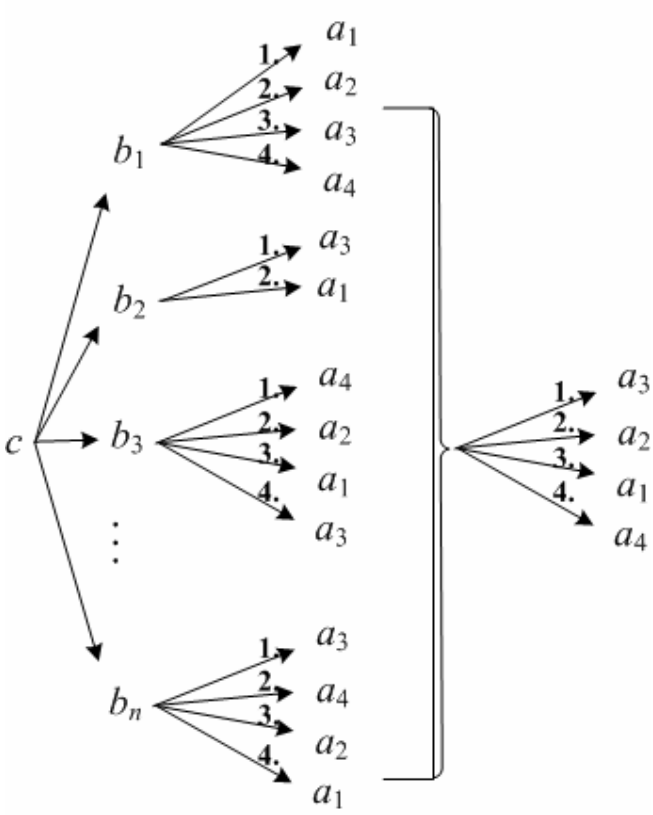


Figure 6: *Open discovery process in LitLinker*.

Both systems, LitLinker and BITOLA, like in Srinivasan and colleagues' case, also use MeSH descriptors as a representation of the MEDLINE documents, instead of using title or abstract words. Here, problems arise since some significant terminology from the subject content of a document may not be covered because MeSH indexers normally use only the most specific vocabulary terms to describe the topic discussed in a document (Nelson et al., 2001).

Recently, Hristovski with colleagues proposed the combined application of two natural language processing systems, BioMedLEE and SemRep (Hristovski et al., 2006) to enhance literature-based discovery. BioMedLEE integrates syntax and semantics. The system uses online biomedical knowledge sources and MedLEE's lexicon (Friedman et al., 2004) developed from clinical documents. The other system, called SemRep (Rindflesch and Fiszman, 2003) is a general knowledge-based semantic interpreter constructed for searching MEDLINE citations and identifying treatments of diseases. It is a symbolic natural language processing system that uses underspecified syntactic analysis and medical domain knowledge from the UMLS Metathesaurus (U.S. National Library of Medicine, 2006a) to identify semantic predications in biomedical text.

Although the majority of the literature-based discovery examples are from the biomedical field, the Swanson's idea can also be explored in other disciplines. Cory described how hidden knowledge within the humanities domains can be extracted from bibliographic databases of the humanities disciplines records (Cory, 1999). He reported that some previously unnoticed analogies were discovered between the epistemological ideas of a nineteenth-century American pragmatic philosopher and an ancient Greek philosopher. Although humanities titles are often imaginative and underdescriptive, Cory showed that logical connections, which were previously unknown and could not be discovered by ordinary search techniques, can be obtained by the application of Swanson's method.

The focus of this thesis is primarily on the areas and methods, where text mining potentially enriches biomedical science and thus interdisciplinary connects information technologies with biomedical expert knowledge. In this respect, we emphasized the specific text mining approaches in real biomedical settings towards extracting knowledge from data. A biomedical area of our particular interest is the research of autism, which is described in more detail in the following chapter.

4 Autism Domain

To better understand the nature of autism spectrum disorders it is necessary to review the scientific literature written about this topic. The purpose of this chapter is therefore to investigate the articles concerning autism, which are available in the MEDLINE database. Our investigation starts with a trend analysis that aims to identify trends in autism scientific research and to compare subsets of autism documents relating to different time periods of their publication. In addition, we review the more recent autism literature in more detail.

4.1 Trend analysis in autism research

Autism is a complex neurodevelopmental disorder. It belongs to a group of pervasive developmental disorders that the fourth revised edition of Diagnostic and Statistical Manual of Mental Disorders categorizes as a group of symptoms of neurological development, associated with early brain mechanisms (American Psychiatric Association, 2000). It is mainly manifested as impairment in social relatedness, communication and as repetitive routines and restricted interests (Georgiades et al., 2007).

For the heterogeneity of this developmental disturbance and its different degrees of affecting children the autistic symptoms occur along a spectrum, more often referred to as autism spectrum disorders. The term Asperger syndrome is often used together with the term autism. There are few content similarities between Asperger syndrome and autism, where no mental retardation is present (Klin and Volkmar, 1995). Both disorders are diagnostically placed within the group of autism spectrum disorders (American Psychiatric Association, 2000).

Throughout the world, the increases in autism rates evidence that autism spectrum disorders are not rare. The estimated prevalence in United States and other developed countries is 5.8 per 1000 children (Hirtz et al., 2007). According to the Autism Society of America, autism is now considered to be an epidemic. The increase in the rate of autism revealed by epidemiological studies and government reports implicates the importance of external or environmental factors that may be changing.

There's also a massive increase of information in the area of autism research, which often impedes to provide a more complete picture of the research results and leads to a fragmented understanding of the nature of autism. The most frequent topics under study include genetics, perception and cognition, neurobiology, physiology and nosology[2] (Matson and LoVullo, 2009).

Despite the number of publications on autism keeps on increasing, the autism domain still lacks a thorough understanding of the underlying phenomena owing to its rather complex nature, and therefore, further investigations are needed (Persico and Bourgeron, 2006). This is the reason why studies seeking for factors that can help to put pieces together into a unique, comprehensible object are so important.

Figure 7 shows the accumulation of the number of autism related articles cited in MEDLINE in the years from 1948 to the end of 2007 in comparison with the growth of all citations in MEDLINE. The data in the figure are the result of an advanced search of the MEDLINE database. An exponential trend of the autism citation growth line clearly demonstrates the increasing acceleration of the autism publications in MEDLINE.

[2] Nosology is the branch of medical science that deals with the classification of diseases.

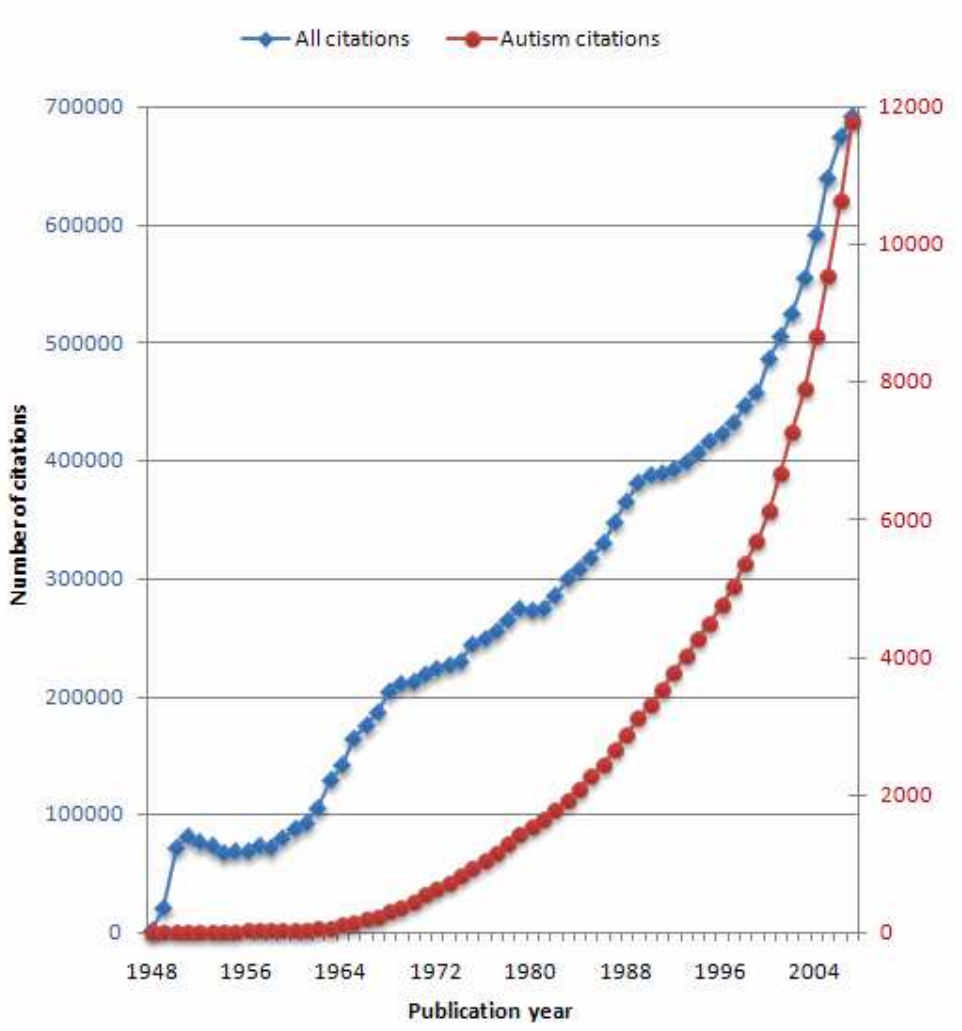


Figure 7: *Growth of autism publications in comparison with the growth of all citations in MEDLINE*. The numbers of autism citations and the numbers of all documents cited in MEDLINE in the period from 1948 to 2007 are presented with two different sets of values for the Y-axis scale.

For trend analysis in autism research we performed the time-dependent analysis of the autism literature from the very first publication in MEDLINE to the autism findings published at the end of 2007. Our literature review is split into five sub-groups according to the periods of publications of autism articles. The first sub-group includes the autism articles that were published since autism was first described by Kanner in 1943 (Kanner, 1943) until the year 1968. The amount of autism research was scarce in that period, in fact this is the smallest of all five groups of documents although it represents the longest period of publication time. From the beginnings of the autism research the emphasis was placed on the supposed difficulties of mothers in relating with their autistic children. In particular, the focus was on the schizophrenic and the so called *refrigerator* mothers[3].

[3] The term *refrigerator mothers* was introduced and popularized by Bettelheim (1967), who blamed parents, usually mothers, for their children's autism. The *refrigerator mothers* were seen as cold and emotionally distant from their children and therefore blamed for their child's disorder despite the lack of empirical evidence to support the theory.

Soon followed the research period, when it was shown that such reasoning on the causes of autism had been false. Several follow-up studies in years 1968-1977 investigated the perception and thinking in autistic children by analysing family home movies, language environment of autistic children and their intellectual characteristics. Many authors in this period raised the question of the early diagnosis within an appropriate nosological system and the suitable methods of treatment of the autistic child.

Publications with more striking findings started to emerge during the late 1970s and 1980s. The genetic studies found associations of autism with the fragile X anomalies and with the Rett syndrome, affecting almost exclusively females. These publications raised the awareness of clinicians about the importance of the systematic genetic screening for the early diagnosis of autism. In this period many authors also reported about the elevated serotonin levels in the blood of autistic children. This way the diversity of studies demonstrated the heterogeneity of the autism aetiology. On the other hand, some consensus was achieved by the autism researchers of the behavioural characteristics associated with the diagnosis of autism.

A substantial research literature was beginning to be published more than forty years after Kanner's groundbreaking work, in the period 1988-1997. Several authors concentrated their attention to speech, imaginative play, and imitation skills. In this period, many papers related to autism also describe studies of features of patients with Tourette's disorder and Asperger syndrome. Asperger syndrome was examined as a subgroup under the autism developmental disorders, which was described for the first time by Asperger in 1944, just a short time after Kanner published his study about infantile autism in 1943 (Klin and Volkmar, 1995). Next, some authors observed Tourette's disorder and suggested that this syndrome may be responsible for some of the genetic heterogeneity in pervasive developmental disorders, such as autism. Besides them, studies included pharmacological treatment with fluoxetine in autism with depression and obsessive-compulsive behaviour.

In more recent studies that were published in the years 1998-2007 marked advances have been made in genetics research. On the other hand, many contradictory studies have been discussing links between autism and immunization with a particular concern regarding the MMR (Measles, Mumps and Rubella) vaccines containing thimerosal. The quantity of research into autism started to increase dramatically in these years; therefore we performed a much detailed review of the autism literature published in this period that we present in the following section.

4.2 Recent autism research

To gain substantial background knowledge about autism, our aim was to review the recent autism literature in some more detail and to identify the most frequent topics researched in this domain in the years 1998-2007. With this intention we retrieved and analysed articles from PubMed Central database that treat problems of autism and were available in full text.

An important goal in our recognition of autism phenomena was to uncover the fundamental concepts that provide necessary clues about autism. To identify some background knowledge from the large amount of digital articles one approach would be to read and manually analyze all available data. Since this is evidently a time consuming task, we instead chose to guide our attention only on the most specific information about the domain of interest. 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 visualization and exploration of large datasets. Here we discuss only briefly the findings of special importance to recent autism research. For further details see Chapter 6.

An interaction of multiple risk factors is considered to contribute to the autistic disorders. However, the etiologies of autism are still largely unknown. Distinctive neuropathological, genetic and environmental studies are of central interest to autism research. In our literature-based study of autism we have mainly focused our attention on biochemical substrates and neurological mechanisms because various neurological and biochemical abnormalities have been identified among individuals with autism. In fact, many mechanisms that might be lying behind neurological disorders of autistic patients have been examined and important advances have been recently made in understanding neural systems that process various types of information. Imaging and other examinations of the autistic brains have shown several brain irregularities, among them the altered neuroanatomy (Bauman and Kemper, 2005), as well as abnormal cellular neurochemistry (DeVito et al., 2007). Within this context, molecular changes in brain development and neurotransmission, morphological distinctions of particular neurons, and regional brain volume abnormalities have frequently been reported (Bethea and Sikich, 2007).

Various neurological conditions in autism reflect also in the heterogeneity of the possible neuropathological causes of this disorder. Therefore, we have started our autism research with an assumption that neurological substances, processes and transformations play a central role in the pathology of autism.

5 Materials and Methods of Data Collection

Materials and the methods of data collection used in our research are presented in this chapter. In particular, we describe the corpus collection that we examined, the XML format version of scientific documents that we utilise to automatically process the documents from MEDLINE, and the Medical Subject Headings classification that serves us for filtering the results returned by the RaJoLink method.

5.1 Research outline

Text mining tools make it possible to discover new knowledge through analysis of text. In order to obtain an improved insight into the autism domain structure and to make valuable new discoveries about autism, we decided to analyse the professional literature about autism that is publicly accessible on the World Wide Web in the MEDLINE database of biomedical publications.

In our experiments that we conducted within the scope of this thesis, we approached as follows:

- We collected scientific articles about autism from the MEDLINE database and in the first place examined the literature on autism by the construction of domain ontologies. To this end, we used OntoGen[4] (Fortuna et al., 2006), the interactive tool for semi-automatic construction of ontologies.

- By text mining the words that rarely appear in the documents about autism we searched for concepts that yield to valuable associations with autism. The search was done in the semi-automated way with the support of the RaJoLink software tool.

- We automatically evaluated the novel candidate targets to elucidate their possible relationships with the autistic spectrum disorders. The automatic evaluation was done with the support of the RaJoLink software tool that retrieved and analysed the documents on candidate targets together with the documents on autism from the MEDLINE database through PubMed searches.

- In the closed discovery process we performed also a more focused investigation of the linking terms between the candidate associative targets and autism by searching for bisociations in the documents' similarity graphs. This investigation was done by concentrating on outlying and their neighbouring documents in the documents' similarity graphs that we constructed on the collection of the autism domain literature together with the documents on an associated domain.

- Finally, we performed experimental evaluation of our method using the literature on migraine. In fact, we demonstrated the RaJoLink's reliability and reproducibility by simulating Swanson's migraine-magnesium discovery, which is considered the gold standard of the literature-based discovery.

---

[4] 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. We used it to construct several autism ontologies on different parts of scientific articles.

5.2 Corpus collection

The material for our experiments was a collection of text documents from MEDLINE database through the PubMed interface (Figure 8). PubMed provides access to the U.S. National Library of Medicine's premiere bibliographic database MEDLINE, as well as to some additional sources of bibliographic information on diseases, public health, pharmacy, pharmacology and other biomedical topics (PubMed, 2008). PubMed automatically retrieves and displays citations on the entered search terms. A single citation may include abstract, full text in PubMed Central or links to full text available elsewhere.

In the MEDLINE database we found 11,969 articles on autism that were published till the end of the year 2007. There were 354 articles with their entire text published in the PubMed Central database, which is the U.S. National Institutes of Health free digital archive of biomedical and life sciences journal articles (PubMed Central, 2006). Other relevant publications were either restricted to abstracts of documents or their entire texts were published in sources outside PubMed.

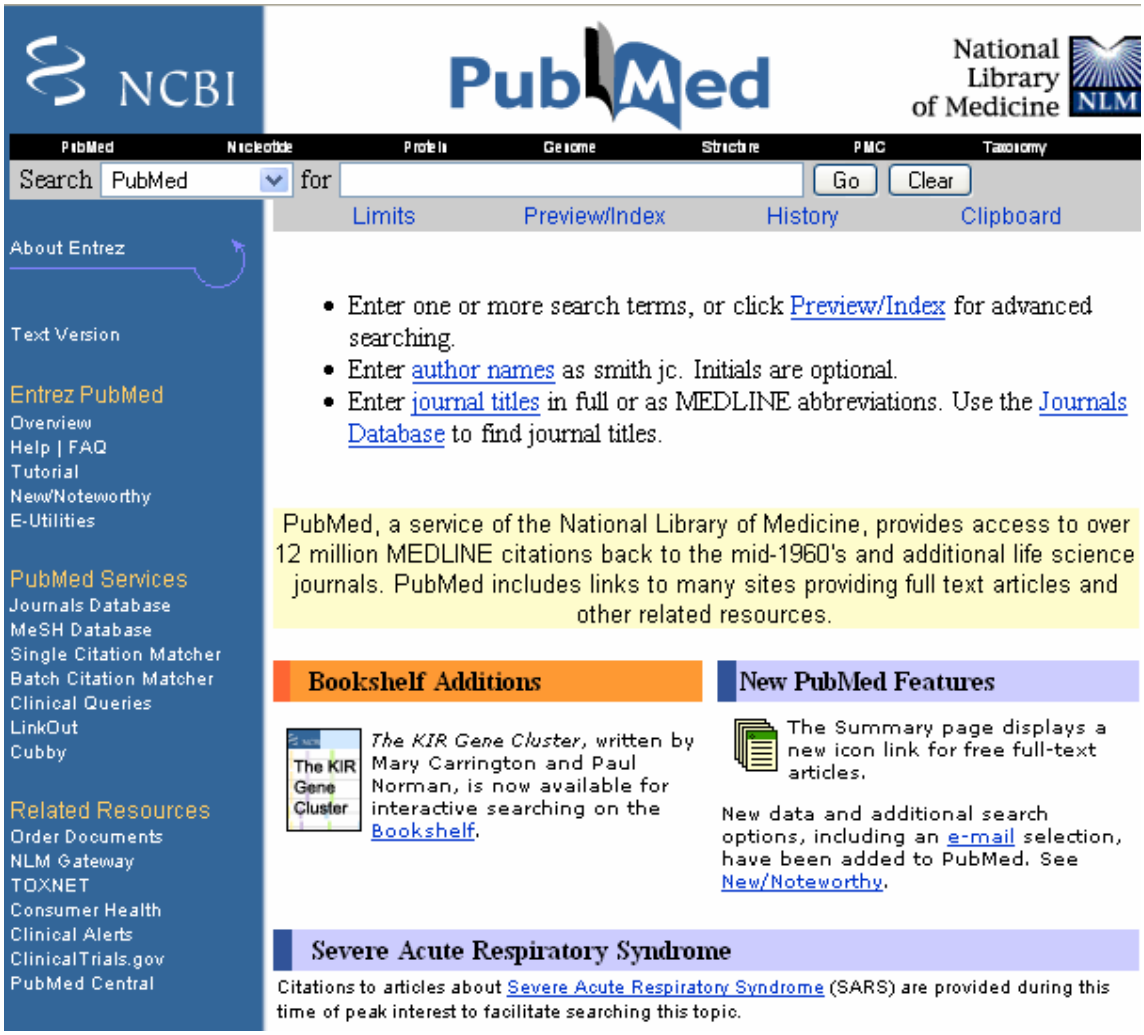


Figure 8: *Screenshot of PubMed that provides access to the articles indexed for MEDLINE*. PubMed is available via the Entrez retrieval system, developed by the U.S. National Center for Biotechnology Information at the National Library of Medicine, located at the U.S. National Institutes of Health.

For the experiments with full texts of articles, we performed preprocessing of such documents by converting the HTML and PDF papers to text, and by deleting graphics, paragraph marks, and manual line breaks from the full text versions, so that each document occupied one record in the input file. Obtaining and handling of full texts of literature requires extra time in terms of locating and converting them into a plain text format. Also, the full-length articles can be only accessed from journals that are available for free or via a particular subscription. However, the full-length biomedical articles contain an abundance of data and if the user could capture important information from them, it is worth spending additional time on obtaining and processing such text.

Unlike titles and abstracts, which are available in HTML or XML format, the full texts from the early issues of some journals are provided only as PDF files. Consequently, their handling requires extra processing time. Therefore, we further restricted the target set of articles from the listed 354 documents to only those that have been published in the last ten years, namely in the period 1998-2007. As a result, we got 214 articles from 1998 forward. For the needs of our experiments and for analysis purposes, we decomposed them to titles, abstracts and bodies of texts. Accordingly, we created three input text files: a file with 214 titles, a file with 214 abstracts and a file with 214 bodies of texts without their respective titles and abstracts.

5.3 Records in XML format

To automatically process the documents from large databases, it is very practical if they are available in XML format. The National Library of Medicine uses also the eXtensible Markup Language (XML) format for disseminating its MEDLINE bibliographic citation data (U.S. National Library of Medicine, 2000).

Each MEDLINE citation has the same structure. The top level element in a citation set is called <MedlineCitation> and contains one entire record. Here, we describe the structure of XML documents and their building blocs as defined by the U.S. National Library of Medicine in the MEDLINE/PubMed document type definitions (U.S. National Library of Medicine, 2005).

Following is a description of the MEDLINE/PubMed elements that we used for literature-based analysis:

- <AbstractText> consists of English-language abstract that is taken rigorously from a published article. If an article does not contain a published abstract, the MEDLINE record lacks the <AbstractText> because the National Library of Medicine does not create one on its own. Typically, the National Library of Medicine does not provide abstracts for articles written before 1975. Whatever the structure of an abstract is, the text is not broken into paragraphs. However, as entry policies for MEDLINE/PubMed citations have changed over the years, some abstracts may be shortened by cutting the end. Such records have informational text enclosed in parentheses (e.g., ABSTRACT TRUNCATED AT 250 WORDS or ABSTRACT TRUNCATED AT 400 WORDS). The present maximum length of abstracts is 10,000 characters for a record;

- <ArticleTitle> contains the complete title of a journal article. The articles that are originally published in a foreign language are translated thus the titles are always provided in English. Some citations from the OLDMEDLINE subset of records are assigned the value *Not Available* for the <ArticleTitle>. These records with missing information were published in the years 1964 and 1965 and cited in the Cumulated Index Medicus. The Cumulated Index Medicus was the first of the medical indexes that provided access to a broad range of biomedical literature in the years from 1960 to 2000 (U.S. National Library of Medicine, 2001).

A sample fragment of a record in the XML format downloaded from the MEDLINE database is presented in Figure 9. It displays the XML tagged format of the MEDLINE citation of our article (Petrič et al., 2009) that was published in the Journal of Biomedical Informatics.

```
<PubmedArticle>
   <MedlineCitation Status="Publisher" Owner="NLM">
      <PMID>18771753</PMID>
      <DateCreated>
         <Year>2008</Year>
         <Month>9</Month>
         <Day>22</Day>
      </DateCreated>
      <Article PubModel="Print-Electronic">
         <Journal>
            <ISSN IssnType="Electronic">1532-0480</ISSN>
            <JournalIssue CitedMedium="Internet">
               <PubDate>
                  <Year>2008</Year>
                  <Month>Aug</Month>
                  <Day>19</Day>
               </PubDate>
            </JournalIssue>
            <Title>Journal of biomedical informatics</Title>
         </Journal>
         <ArticleTitle>Literature mining method RaJoLink for uncovering relations
         between biomedical concepts.</ArticleTitle>
         <Pagination>
            <MedlinePgn/>
         </Pagination>
         <Abstract>
            <AbstractText>To support biomedical experts in their knowledge
            discovery process, we have developed a literature mining method called
            RaJoLink for identification of relations between biomedical concepts
            in disconnected sets of articles. The method implements Swanson's ABC
            model approach for generating hypotheses in a new way. The main
            novelty is a semi-automated suggestion of candidates for agents a that
            might be logically connected with a given phenomenon c under
            investigation. The choice of candidates for a is based on rare terms
            identified in the literature on c. As rare terms are not part of the
            typical range of information, which describe the phenomenon under
            investigation, such information might be considered as unusual
            observations about the phenomenon c. If literatures on these rare
            terms have an interesting term in common, this joint term is declared
            as a candidate for a. Linking terms b between literature on a and
            literature on c are then searched for in the closed discovery to
            provide additional supportive evidence for uncovered connections. We
            have applied the method to the literature on autism and have used
            MEDLINE as a source of data. Expert evaluation has confirmed that the
            discovered relations might contribute to a better understanding of
            autism.</AbstractText>
         </Abstract>
```

Figure 9: *The sample record from the MEDLINE database in the XML format*. The displayed fields of the sample record are part of our article from the Journal of Biomedical Informatics, viewed as MEDLINE citation in XML format (Source: U.S. National Library of Medicine, September 2008).

For automatic access to the MEDLINE data, which has to be performed outside of the regular web query interface, called Entrez, we used the ESearch tool of the Entrez Programming Utilities (Sayers and Wheeler, 2004). In particular, we operated with the following utility parameters:

- *Database name* (*DbName*), where the PubMed database is the default value of the parameter *DbName*;

- *Date Ranges* to limit query results bounded by two specific dates, namely the *mindate* and the *maxdate* parameters;

- *Date Type* that limits dates to a specific date field in a database. Actually, we use the *edat* type of dates, which limits query results according to the date when a citation was added to PubMed.

The Entrez Programming Utilities are capable of retrieving data records that already exist inside the Entrez system. Among them, the ESearch tool responds to a text query and returns data corresponding to the results of the query submitted to the Entrez system. As the results from ESearch are maintained in the user's environment, the maximum number of retrieved records (URL parameter *retmax*) is 10,000 records for a search. If there is a need to retrieve more than 10,000 records from MEDLINE we suggest running a combination of queries with different date ranges.

5.4 MeSH classification

The Medical Subject Headings (MeSH) thesaurus is produced by the U.S. National Library of Medicine, which has also been maintaining it since 1960. The formal definition of the goal of MeSH is "to provide a reproducible partition of concepts relevant to biomedicine for purposes of organization of medical knowledge and information" (Nelson et al., 2001). MeSH terms form a controlled vocabulary that is primarily used for subject indexing and searching of journal articles in MEDLINE database and other catalogues of the U.S. National Library of Medicine.

The MeSH structure consists of three major components: the headings, the subheadings and the supplementary concept records (Nelson et al., 2001). The subheadings that are also known as qualifiers are assigned to the main headings by indexers and are mostly used as the topical qualifiers. Supplementary concept records are mainly related to chemicals and drugs and include names of substances, synonyms, structural chemical names, registry numbers, and other notes that are daily added to the Mesh. Main headings within the MeSH classification are used as indexing terms in the MEDLINE database to denote the major topics that are discussed by the cited article. On the other hand, the MeSH thesaurus represents descriptors that reflect the broad meaning of the term under observation (Figure 10).

MeSH Heading	Autistic Disorder
Tree Number	F03.550.325.125
Scope Note	A disorder beginning in childhood. It is marked by the presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activity and interest. Manifestations of the disorder vary greatly depending on the developmental level and chronological age of the individual. (DSM-IV)
Entry Term	Autism
Entry Term	Autism, Early Infantile
Entry Term	Autism, Infantile
Entry Term	Kanner's Syndrome
Allowable Qualifiers	BL CF CI CL CO DH DI DT EC EH EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI SU TH UR US VI
Previous Indexing	Autism (1966-1970)
Previous Indexing	Schizophrenia, Childhood (1966-1970)
Online Note	use AUTISM, INFANTILE to search AUTISM, EARLY INFANTILE 1971-80; use SCHIZOPHRENIA, CHILDHOOD 1968-70
History Note	1981(1966)
Date of Entry	19990101
Unique ID	D001321

Figure 10: *MeSH descriptor data for autistic disorder*. (Source: U.S. National Library of Medicine, July 2008).

The basic building block of the MeSH thesaurus, which is the descriptor class, is formed by one or more terms that represent a concept. The descriptors in the MeSH thesaurus are hierarchically related by parent-child relationships so that each descriptor has at least one parent. Besides, the relationships among concepts can be explicitly represented as relationships within a descriptor class (Nelson et al., 2001).

From these points of view the MeSH tree structures that form the hierarchical MeSH thesaurus can be regarded as an ontology. In the entire structure of the MeSH thesaurus, one particular term can appear in different concepts. For example, the term *knowledge* can be found under two MeSH categories, namely under the MeSH category Humanities – K01, labelled with the MeSH tree number K01.468 and under the MeSH category Information Science – L01, labelled with the MeSH tree number L01.535.

The MeSH terminology is organized into the eleven-level hierarchical structure. The broad medical headings are situated at the most general levels of the hierarchy, while the more specific medical headings are at narrower levels. The top-level categories in the MeSH hierarchy are: Anatomy - A, Organisms - B, Diseases - C, Chemicals and Drugs - D, Analytical, Diagnostic and Therapeutic Techniques and Equipment - E, Psychiatry and Psychology - F, Biological Sciences - G, Natural Sciences - H, Anthropology, Education, Sociology and Social Phenomena - I, Technology, Industry, Agriculture - J, Humanities - K, Information Science - L, Named Groups - M, Health Care - N, Publication Characteristics - V, Geographicals - Z.

As an example, Figure 11 presents the hierarchy of the MeSH tree as we navigate from the Mental Disorders category at the tree top to the Autistic Disorder category near the bottom of the MeSH tree structures.

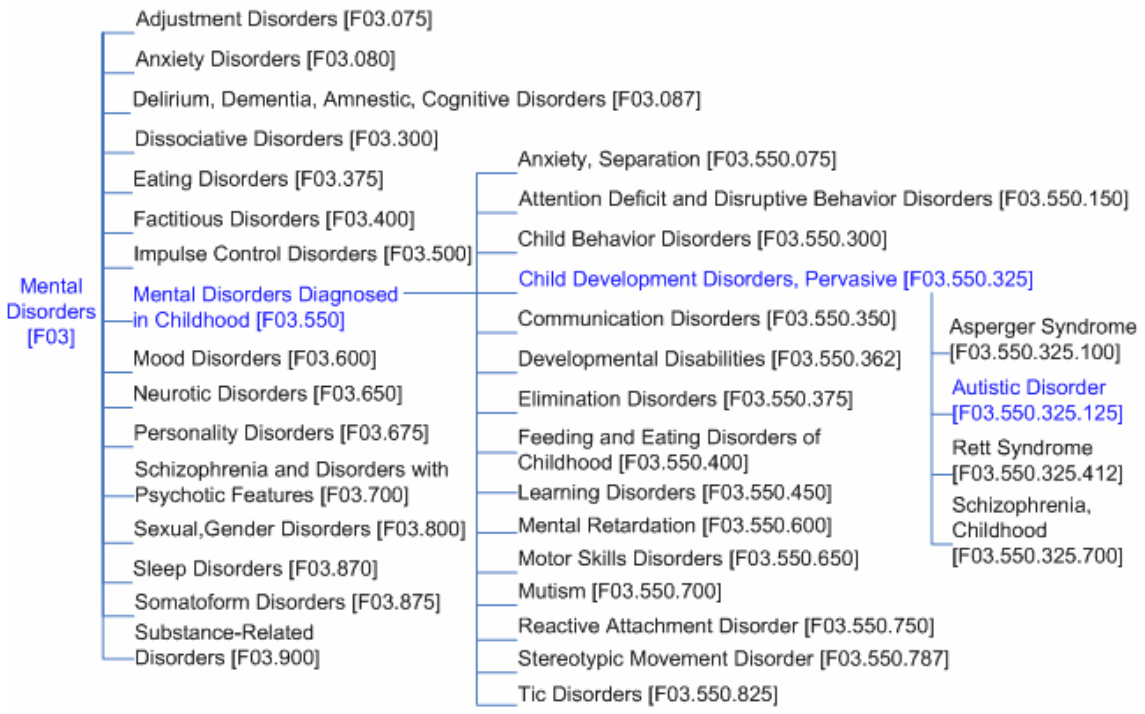


Figure 11: *MeSH tree structure for the Mental Disorders category [F03] when approaching to the Autistic Disorder category [F03.550.325.125]*. (Source: U.S. National Library of Medicine, September 2008).

We take advantage of the MeSH classification to map the terms from free text to the concepts within this biomedical controlled vocabulary. We evaluate against the MeSH thesaurus each string variant that results from the text pre-processing and the morphological analysis of words. This way we facilitate the filtering out of terms by providing the possibility to select only those categories from the MeSH tree structure that the user is interested in.

6 Structuring Domain Knowledge with Ontology Construction

To obtain an overview of the fundamental concepts of autism domain knowledge we constructed several domain ontologies on literature about autism and performed analyses of the constructed ontologies. In this chapter we provide details about our experiments in the semi-automated ontology construction in autism domain. In this objective, we investigated how separate parts of scientific articles, such as titles, abstracts and full texts, influence the constructed ontology.

6.1 Ontologies

For successful text mining a wide background knowledge concerning the problem domain presents a substantial advantage. Ontologies in general with their capability to share a common understanding of domains support researches with the ability to reason over and to analyze the information at issue (Joshi and Undercoffer, 2004). In information science, ontology is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them. In this manner, domain ontologies represent background information from the application domains. 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.

Throughout each period of science, ontologies have been used as a means to organize scientific information and, more importantly, to provide a common vocabulary of concepts. From this perspective, ontologies are part of the common-sense understanding of the world, which define the concepts and structures in a domain.

Ontologies are used in information science as a form of knowledge representation of the world or some part of it. In general, ontologies include descriptions of objects, concepts, attributes and relations between objects. They integrate and conceptualize the heterogeneity of the domain terminologies that can be identified in text. Therefore, ontologies reflect the content and the structure of the knowledge as it can be recognized through the use of terms in the inspected literature. The literature that is utilised in the construction of topic ontologies must be carefully selected before it is processed and considered for analyses.

Until recently, the practice of ontology construction has relied mostly on the manual extraction of interesting concepts from scientific literature and their organisation in a suitable hierarchy. Nowadays, the largely increased amount of scientific publications requires automated support for such a task. With new knowledge technologies, selected scientific articles can be processed semi-automatically, and therefore, the process of ontology construction can be made more effective and feasible in practice. Thus, ontologies are particularly important when the process of knowledge acquisition embraces insight and understanding of a specific domain.

Ontologies as those illustrated in Figures 12 and 13 actually helped us to review and understand the complex and heterogeneous spectre of scientific articles about autism. In fact, the examination of the autism literature indicates a wide diversity in the autism studies. In order to broadly assess the changes in the autism research over time, we initially performed a trend analysis. Trend analysis can be used in text mining to recognize the trends in journal papers across multiple document subsets over time (Feldman and Sanger, 2006). We performed our domain research by semi-automatic construction of ontologies with the computational support of OntoGen (Fortuna et al., 2006).

The ontology-based trend analysis of the autism literature that is represented by the ontology in Figure 12 is provided on the bases of the collection of autism documents that we analysed and by relationships between concepts that these documents address. This way the construction of the domain ontologies enables the observation of the general trends of the documents topics across different publication periods.

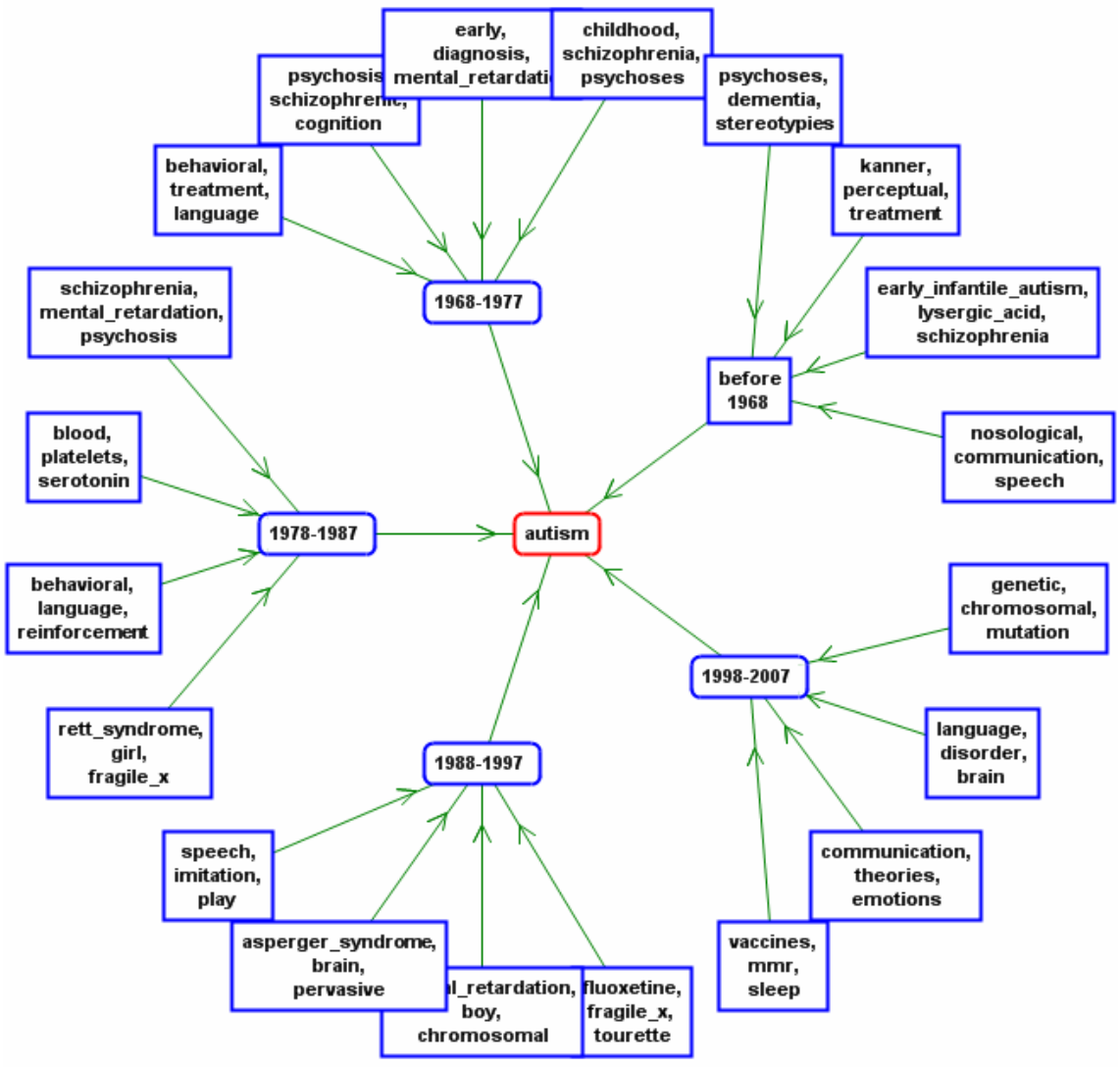


Figure 12: *Autism trend analysis*. Trend of autism research is represented by the ontology, where documents were divided into 5 periods according to the year of their publication.

For the ontology-based trend analysis we divided autism documents from the MEDLINE database into five time-dependent datasets by considering the publication date of the documents. Each dataset, except the first one (i.e. for the period before 1968) comprised the autism research that falls within a period of 10 years as follows: 1968-1977, 1978-1987, 1988-1997, and 1998-2007. Due to the smaller number of MEDLINE citations we included all the early publications on autism before year 1968 within a single dataset. The five datasets consisted of abstracts of the autism documents and of the documents' titles for all those articles that don't have their abstracts available in MEDLINE.

We split each of the concepts on the first level of ontology into sub-concepts and named them with the most informative keywords that were automatically extracted from the concepts documents from each period. From such representations we were able to detect those topics that occurred during several periods of time (e.g., behavioural, language, communication, fragile X, and other chromosomal studies of autism) and can be therefore regarded as the most intensively researched fields of autism.

We subsequently built the autism domain ontologies on autism articles from PubMed Central database that were available in full text and published in the years 1998-2007. An example of such ontology in the form of a tree-based concept hierarchy is given in Figure 13. In this example, we split the first level of this concept hierarchy into four concepts. We further split each of these concepts into at least six sub-concepts.

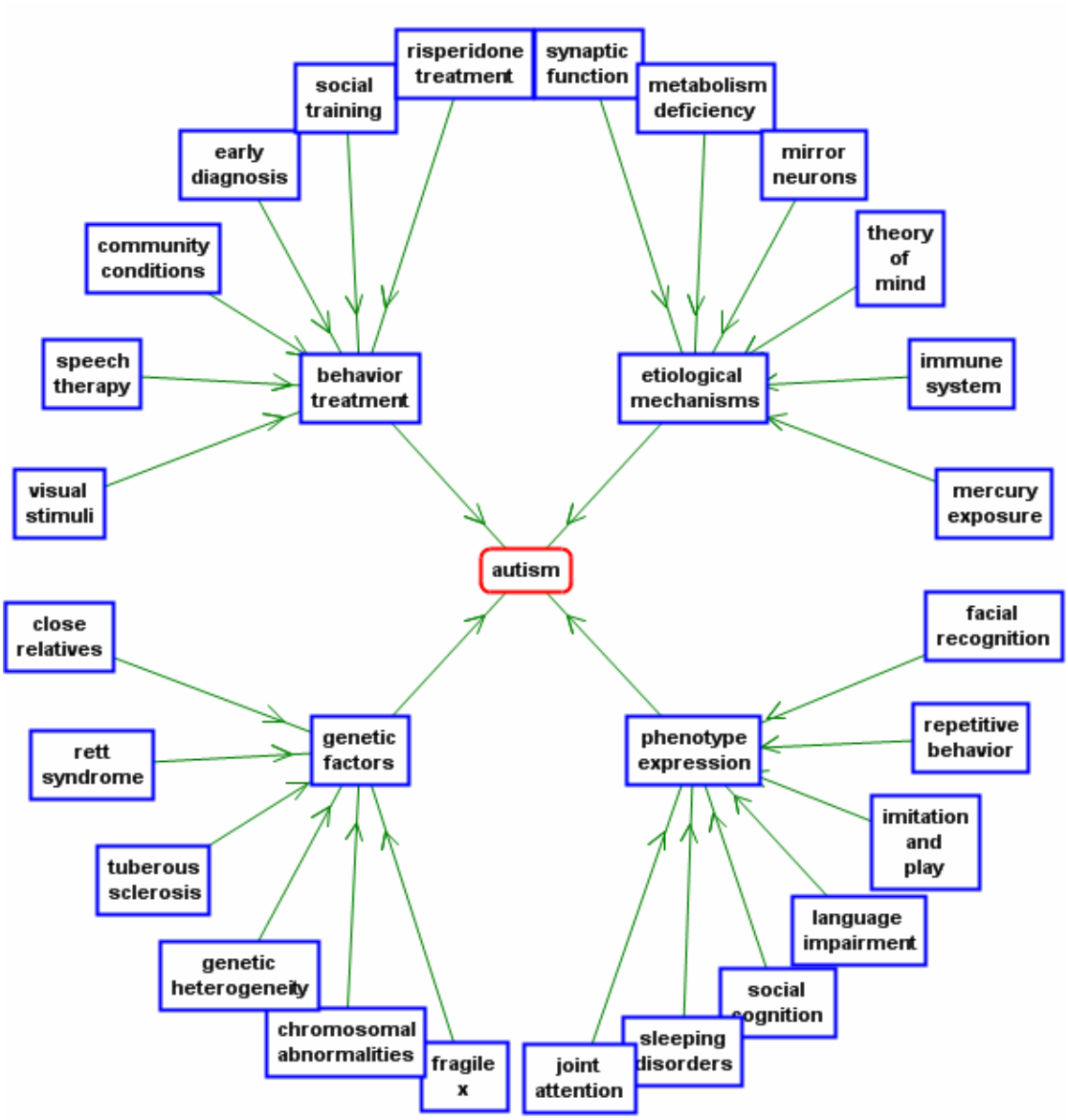


Figure 13: *A two-level autism ontology*. Concepts are renamed according to autism survey literature, based on the keywords suggested by OntoGen.

By constructing ontologies we extracted key information and knowledge from the scientific literature and got insights into the structure of large literature collections. The medical expert that evaluated ontologies, which were constructed on the autism literature found a tree-based view on the ontologies to be very intuitive representation of the autism scientific research. Such visualization through the semantic analyses of scientific literature helped us exploring the autism domain and understanding its heterogeneous nature. Therefore, we integrated ontologies in our methodology because they model a domain under research in a way that facilitates understanding. In this manner ontologies complement the knowledge discovery process.

6.2 Semi-automatic ontology construction

Traditionally, ontologies for a given domain are constructed manually using some sort of language or representation and rely on the manual extraction of common-sense knowledge from various sources.

Recently, several programs that support manual ontology construction have been developed, such as Protégé (Gennari et al., 2002). Since manual ontology construction is a complex and demanding process, there is a strong tendency to provide a computerised support for the task. Based on text mining techniques that have already proven successful for the task, OntoGen (Fortuna et al., 2006) is a tool that enables the interactive construction of ontologies from text documents in a selected domain. A user can create concepts, organise them into topics and also assign documents to concepts. A screenshot of OntoGen version 2.0.0.0 is shown in Figure 14.

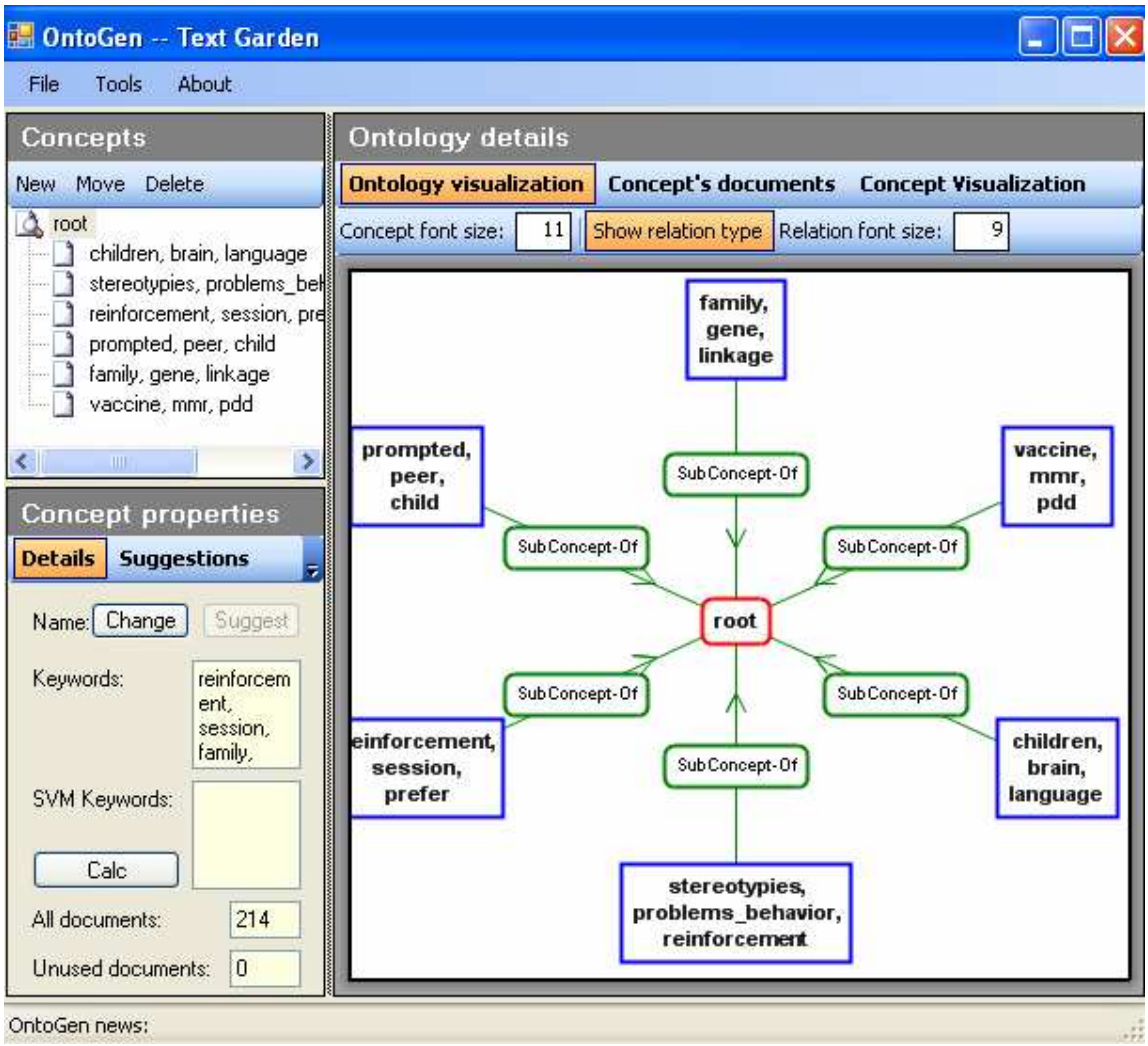


Figure 14: *Screenshot of OntoGen, version 2.0.0.0.* A tool for interactive topic ontology construction (Fortuna et al., 2006).

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 the automatic assignment of documents to the concepts (Fortuna et al., 2006). Our main motivation for using OntoGen was to gain a quick insight into a given domain by semi-automatically generating the main ontology concepts from the domain's documents. The semi-automatic ontology construction method implemented in OntoGen incorporates basic text mining principles. The input for the tool is a collection of text documents. Documents are represented as vectors, which together are often referred to as a vector space model. Using this representation, similarities between two documents can be defined as the cosine of the angles between the two corresponding vector representations. When suggesting new concepts, OntoGen uses a KMeans clustering technique (Jain et al., 1999) and a keyword extraction method (Brank et al., 2002).

6.3 Experiments on documents about autism

We investigated 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 studied the differences in automatically constructed ontologies from titles, abstracts and bodies of texts respectively. While some experts suggest that the more text one can obtain, the better the constructed ontology (Liu et al., 2003), others advocate a more systematic approach that relies on comparably balanced parts of explored texts (Cohen et al., 2005).With the experiments we aimed to clarify this dilemma. Thus, our main motivation was to analyse how separate parts of scientific articles influence the constructed ontologies. In this comparison, we decided to take into account only the top-level ontology concepts, mostly because comparing full-scale ontologies can become a very intricate task (Brank et al., 2005).

Initial results presented in our early study (Petrič et al., 2006a; Cestnik et al., 2007) encouraged further investigation that enabled us to present our findings in a more systematic fashion. When evaluating which parts of articles would be more appropriate for ontology construction, we assessed two criteria: first, the pair-wise similarity of the constructed ontology concepts, and second, their resemblance to the commonly accepted concepts in a given domain.

When designing the experiments, we had two goals in mind. First, we wanted to become acquainted with the domain in the sense that we understand better the underlying concepts. Second, we wanted to evaluate various ontologies constructed on various parts of documents, such as titles, abstracts and texts. In addition, we also tried to evaluate the content compliance between titles, abstracts and entire bodies of texts of the related documents.

Finally, we also wanted to experiment with various values of the parameter *k* used by OntoGen's *K*-means clustering algorithm. Clustering algorithms, such as K-means clustering, are useful tools for data mining; however, when we have to cluster datasets, it is not always clear which is the most appropriate number of clusters (parameter *k*) to use (Jain et al., 1999). The earlier version of OntoGen automatically proposed the use of eight clusters as a default. However, it is strongly recommended to experiment also with various other values of *k* in order to determine the best result for the domain under investigation.

The ontologies were built with two values for the parameter *k*: first, with *k*=8, which was automatically suggested by the earliest version of OntoGen, and second, with *k*=5, which experimentally turned out to be a well-balanced trade off between complexity and comprehensibility in this domain. The resultant ontology concepts are illustrated by the example ontology in Figure 15.

Moreover, the results obtained with *k*=5 were more in accordance with the concepts found in the autism survey literature (Zerhouni, 2004) and were also confirmed by an expert in the autism domain. In this way, OntoGen generated eight and five concepts respectively on the first level of domain ontology for each of the input files (titles, abstracts and bodies of texts). Each concept was described with the three most relevant keywords as suggested by OntoGen.

For these experiments we used three input text files: a file with 214 titles, a file with 214 abstracts and a file with 214 bodies of texts on autism that we obtained by our search in the PubMed Central database. Each text file was used separately as an input for OntoGen; in the process of semi-automatic ontology construction, we used OntoGen to construct several top-level ontology concepts and describe them with suggested keywords.

Our evaluation of the obtained ontology concepts was first performed at vocabulary level by comparing keywords of various concepts and analysing the sets of documents that corresponded to each concept. Next, concept descriptions were presented to the medical expert, who also evaluated the concepts from her perspective.

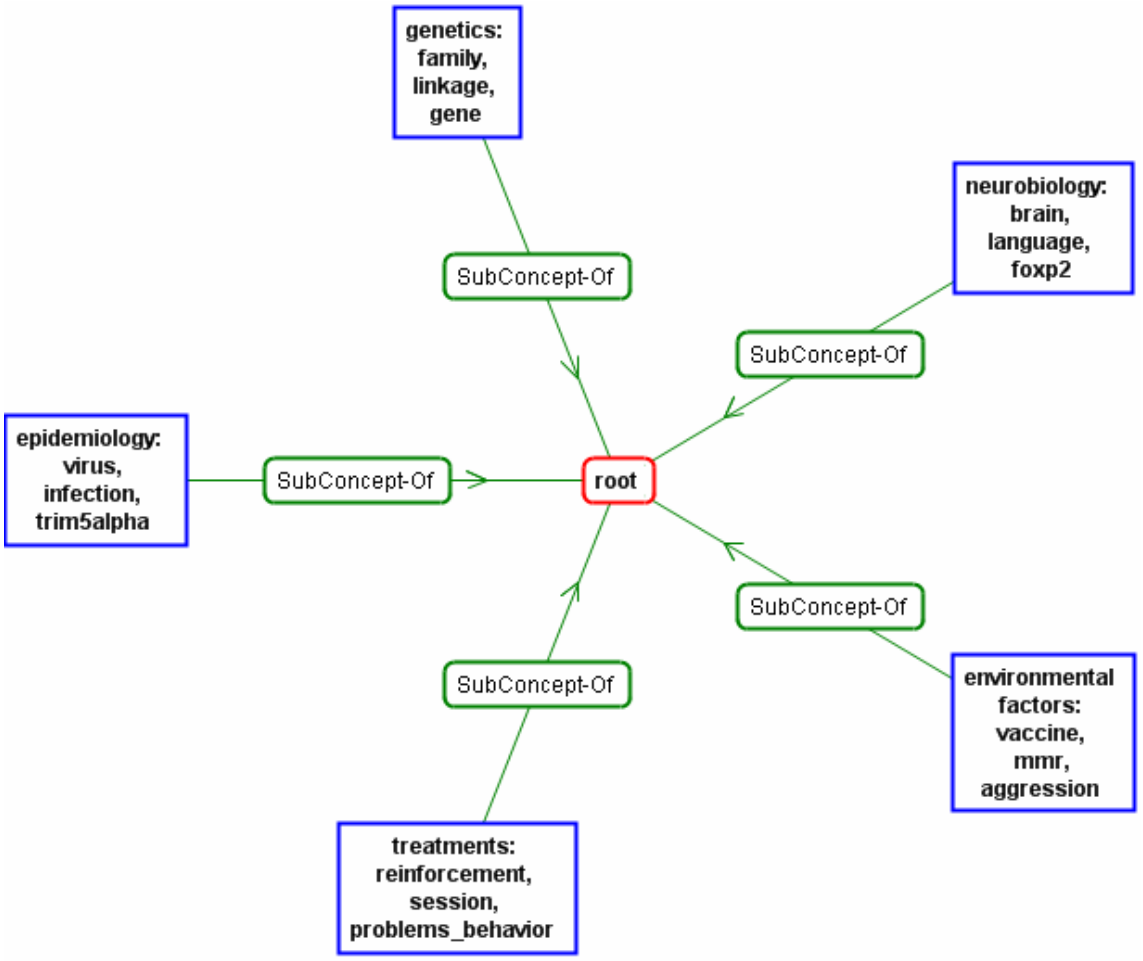


Figure 15: *Top-level autism ontology concepts*. Original concepts descriptions (three for each concept) as suggested by OntoGen are included for easier identification.

6.4 Experimental results of ontology construction in the autism domain

Our comparison of ontologies, built with the tool OntoGen, was made on 214 articles that treat problems of autism and are available in the PubMed Central database. When dealing with complex phenomena, a good strategy is to decompose it to more manageable parts (Zupan et al., 1997). The document corpora can usually be divided by hierarchical structure of a document into logical sections such as title, abstract and main body (Hollingsworth et al., 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 phenomenon, our goal was also to review the autism literature and to identify the most frequent topics researched in this domain.

Using OntoGen we displayed sub-concepts of the 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 (Fortuna et al., 2006). In fact, OntoGen implements two keyword extraction techniques. The first one results in keywords extracted from the concept's centroid vector. The second one extracts keywords from the concept's Support Vector Machine linear model by dividing documents within the concept from the neighbouring documents and thus takes into account the context of the topic. However, when describing concepts of autism ontologies we used the first method (i.e. the concept's centroid vector), because we didn't know the contexts of the autism domain in advance and

we therefore were basically focused on finding the most important words within the concepts that served as keywords. 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.

Finally, we compared the locations of a particular document within different ontologies and performed an analysis on how the variation of parameter *k* impacts the positioning of the title, abstract or full text of that particular document. When parameter *k* was set to 5 the title, the abstract and the body of a given document frequently appeared in the lexically related clusters with similar or even equal keywords in their names. However, when parameter *k* was set to 8, this correlation was not significant. This observation suggests that different number of sub-concepts affects also the positioning of a particular document's parts in a semantic space.

Tables 1 to 6 present the results of our experiments on titles, abstracts and full texts of 214 articles on autism. Each table from Table 1 to Table 6 contains ontology concepts described using three keywords and the number of related documents.

Table 1: *Eight concepts of autism ontology generated from 214 titles*.

ID	Keywords	Number of documents
0	root	214
1	genes, susceptibility, specific	32
2	disorders, linkage, case	32
3	preference, assessment, affects	31
4	reinforcement, children_autism, early	27
5	functioning, syndrome, analysis	26
6	autism, teach, child	25
7	vaccination, schedules, activated	24
8	social, evidence, chromosome	17

Table 2: *Eight concepts of autism ontology generated from 214 abstracts*.

ID	Keywords	Number of documents
0	root	214
1	gene, linkage, regional	60
2	reinforcers, preferred, stimulus	41
3	language, age, children	28
4	stereotypy, behavioural, problems_behavioral	26
5	teach, question, procedure	18
6	vaccine, mmr, mmr_vaccine	17
7	parent, mmr, vaccine	16
8	sensory, sounds, auditory	8

Table 3: *Eight concepts of autism ontology generated from 214 bodies of texts*.

ID	Keywords	Number of documents
0	root	214
1	linkage, family, gene	55
2	reinforcement, session, aggression	38
3	stereotypes, reinforcement, problems_behavior	27
4	executive, nv[5], cortical	26
5	vaccine, mmr, mmr_vaccine	25
6	prompted, script, teaching	21
7	chemical, infant, sleep	14
8	ht, secretin, legs	8

[5] Non-verbal

The evaluation of the obtained results show differences between ontology concepts constructed from titles, abstracts, and related bodies of texts. Figures 16 and 17 compare the distribution of titles, abstracts and entire bodies of texts when documents were divided into eight and then into five sub-concepts within each ontology.

Each document can only be mapped to one concept. However, different keywords may appear in a concept description according to different approach to constructing ontologies. OntoGen automatically generates three keywords, which constitute the semantic description of documents that are mapped to a particular concept. In our experimental study, the distribution of documents among eight concepts of the title ontology (Table 1) is rather uniform. In contrast, the ontologies of eight abstract concepts (Table 2) and eight text concepts (Table 3) both show one major sub-concept of documents that treat genetics and another important group that describes reinforcers or stimuli for autistic patients (Figure 16).

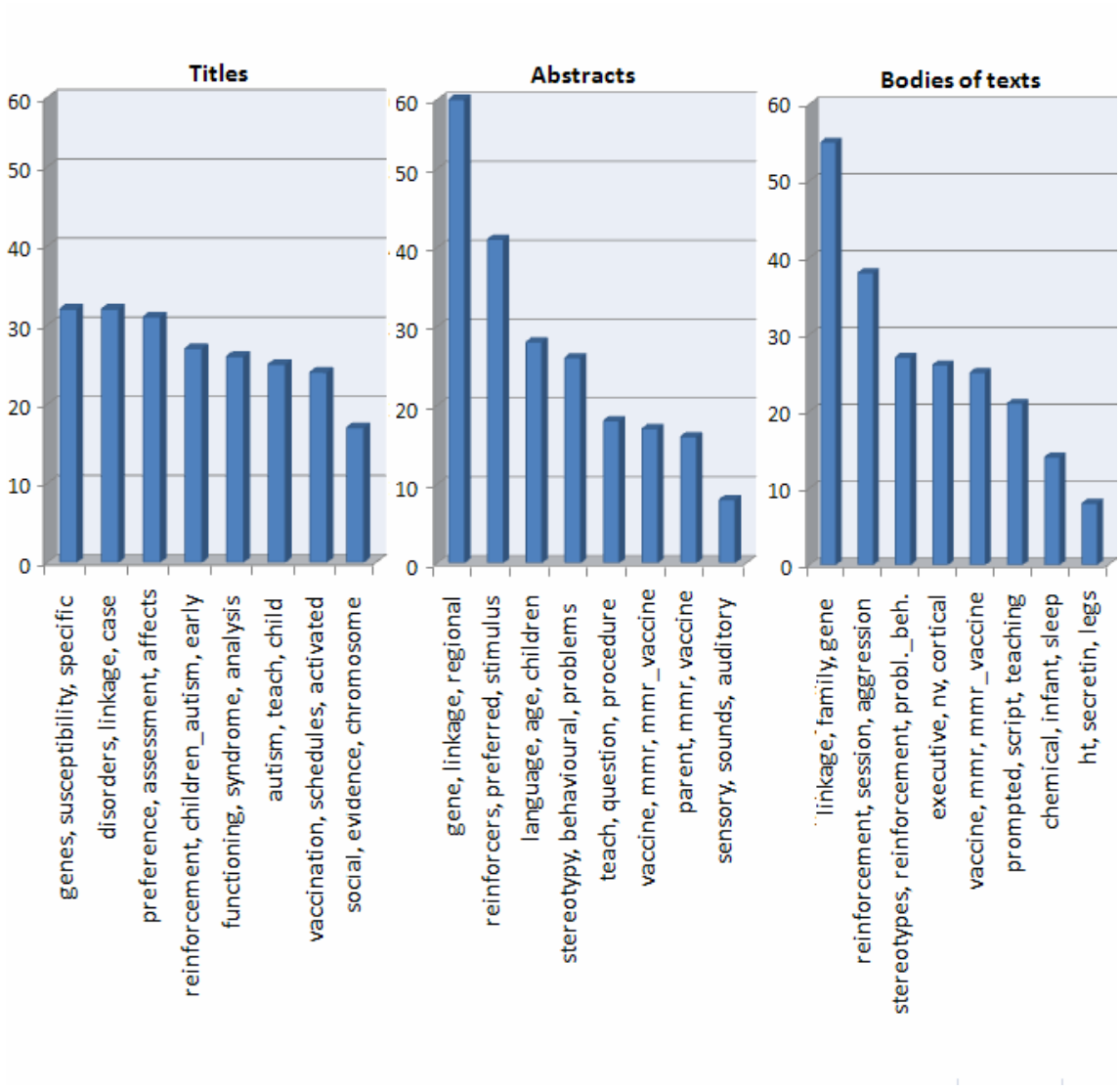


Figure 16: *Comparison between the distributions of documents when they were divided into 8 sub-concepts.*

After experimenting with OntoGen's default parameter, k=8, we also constructed top-level ontology concepts with several other values for k, ranging from 2 to 15. As a result, we discovered that value 5 for k represents a well-balanced trade off between the complexity and comprehensibility of the single-level ontology concepts in this domain. Although the concepts generated with other values of k also revealed

some interesting domain properties, they were either too broad when *k* was small or too narrow when *k* was large. Therefore, a careful selection of value of *k* is a very important prerequisite when constructing ontologies in a semi-automatic way.

Table 4: *Five concepts of autism ontology generated from 214 titles*.

ID	Keywords	Number of documents
0	root	214
1	autism, children_autism, children	67
2	genetic, chromosome, linkage	50
3	disorders, spectrum, neurodevelopmental	39
4	reinforcement, effects, behavior	39
5	syndrome, detection, social	19

Table 5: *Five concepts of autism ontology generated from 214 abstracts*.

ID	Keywords	Number of documents
0	root	214
1	linkage, gene, regional	55
2	language, foxp2, children	52
3	reinforcers, behavioural, problems_behavioral	49
4	reinforcers, vaccine, aggression	46
5	virus, infection, trim5alpha	12

Table 6: *Five concepts of autism ontology generated from 214 bodies of texts.*

ID	Keywords	Number of documents
0	root	214
1	reinforcement, session, trial	72
2	linkage, family, gene	71
3	reinforcement, sleep, infant	37
4	vaccine, mmr, mmr_vaccine	24
5	infection, pml, patients	10

Document distributions in ontologies of five sub-concepts are a little different (Figure 17). There are two major groups of titles (Table 4) and bodies of texts (Table 6). The largest group of titles describes autism in general, whereas the largest text group relates to reinforcement trials. The second major group in both cases (titles and texts) deals with genetics. The distributions of abstracts (Table 5), in contrast, shows two very important groups that both treat differing aspects of genetics. While the first major group of abstracts is described using clear genetic keywords, the second major group of abstracts includes, among others, keyword *foxp2*, which is a gene important for the development of speech.

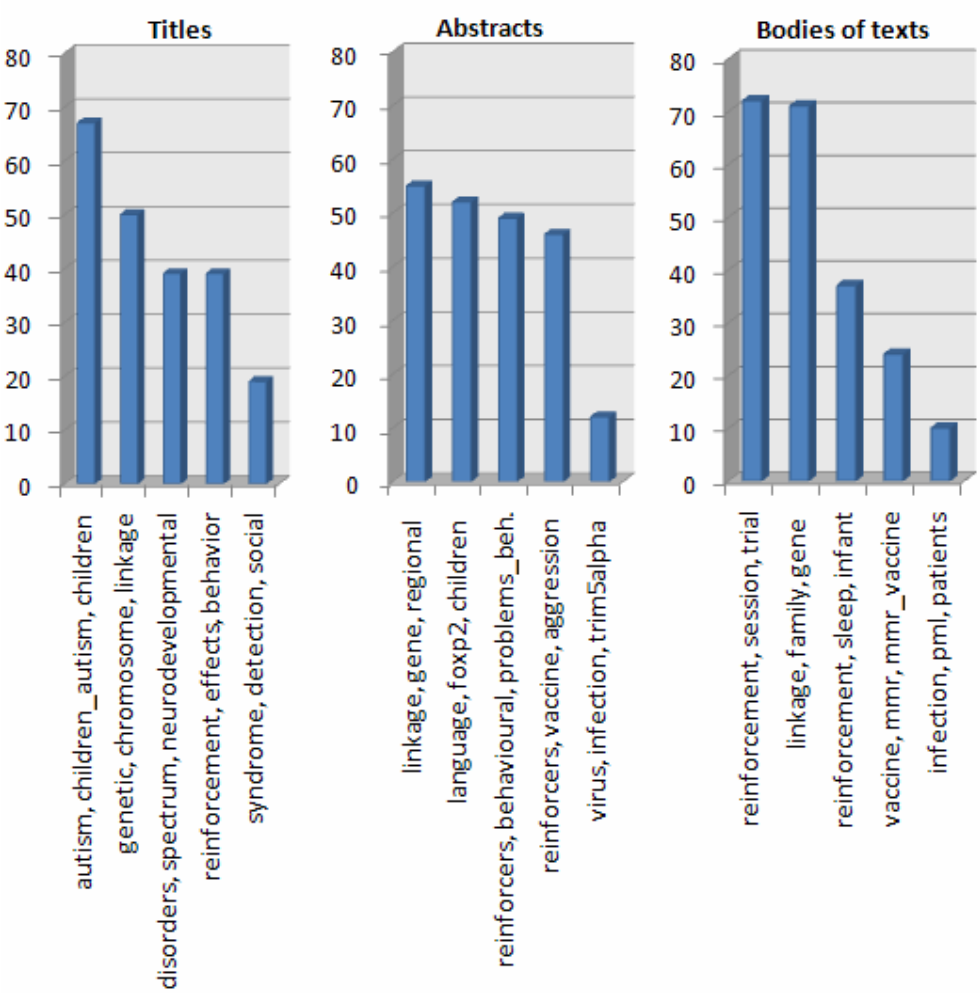


Figure 17: *Comparison between the distributions of documents when they were divided into 5 sub-concepts.*

6.5 Evaluation of experimental results

With the purpose to perform the comparison of ontologies and to describe the similarity between them, Maedche and Staab propose evaluations at two different levels, the lexical and the conceptual (Maedche and Staab, 2002). Firstly, they recommend investigating how terms are used at the vocabulary level to determine the meaning of text. Secondly, they propose to study the conceptual relations between the terms, where they evaluate ontologies by comparing taxonomies of ontologies and by comparing relations in ontologies.

In most cases, ontologies are rather complex structures. It is therefore often more reasonable to focus the attention on the evaluation of separate levels of ontology, rather than on the direct evaluation of whole ontologies (Brank et al., 2005). In our comparison of the ontology concepts from autism, built using OntoGen, we focused at the vocabulary level of the obtained concept descriptions and related concept documents. We observed the distribution of documents within individual ontology groups on the first level of each ontology model (first-level sub-concepts of autism domain), considering terminology that was selected by OntoGen for the presentation of concepts.

From the comparison between the ontology of 8 texts groups versus 8 abstracts groups (Figure 18), the major similarity is shown between the groups 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.

The comparison of ontologies with five groups of texts and five 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 lexical similarities between the ontology of abstracts and the ontology of titles.

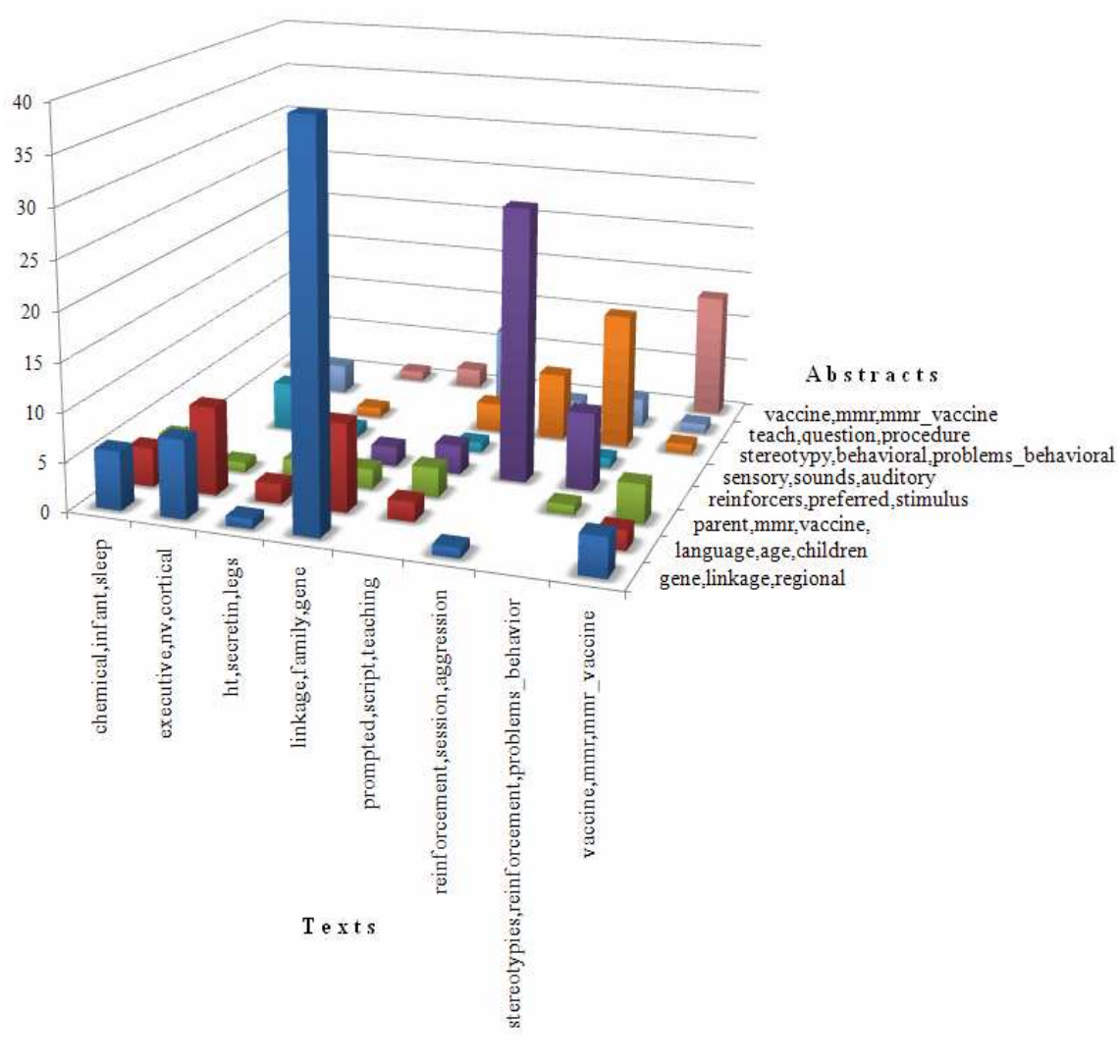


Figure 18: *Comparison between the distributions of documents belonging to the ontology concepts of abstracts and bodies of texts when documents were divided into 8 sub-concepts.*

Among the groups of documents which belong to the certain of five subgroups of texts and at the same time to its relative subgroup of abstracts, the largest similarity is between the groups of genetic texts and abstracts (Figure 19), which cover the same 51 documents.

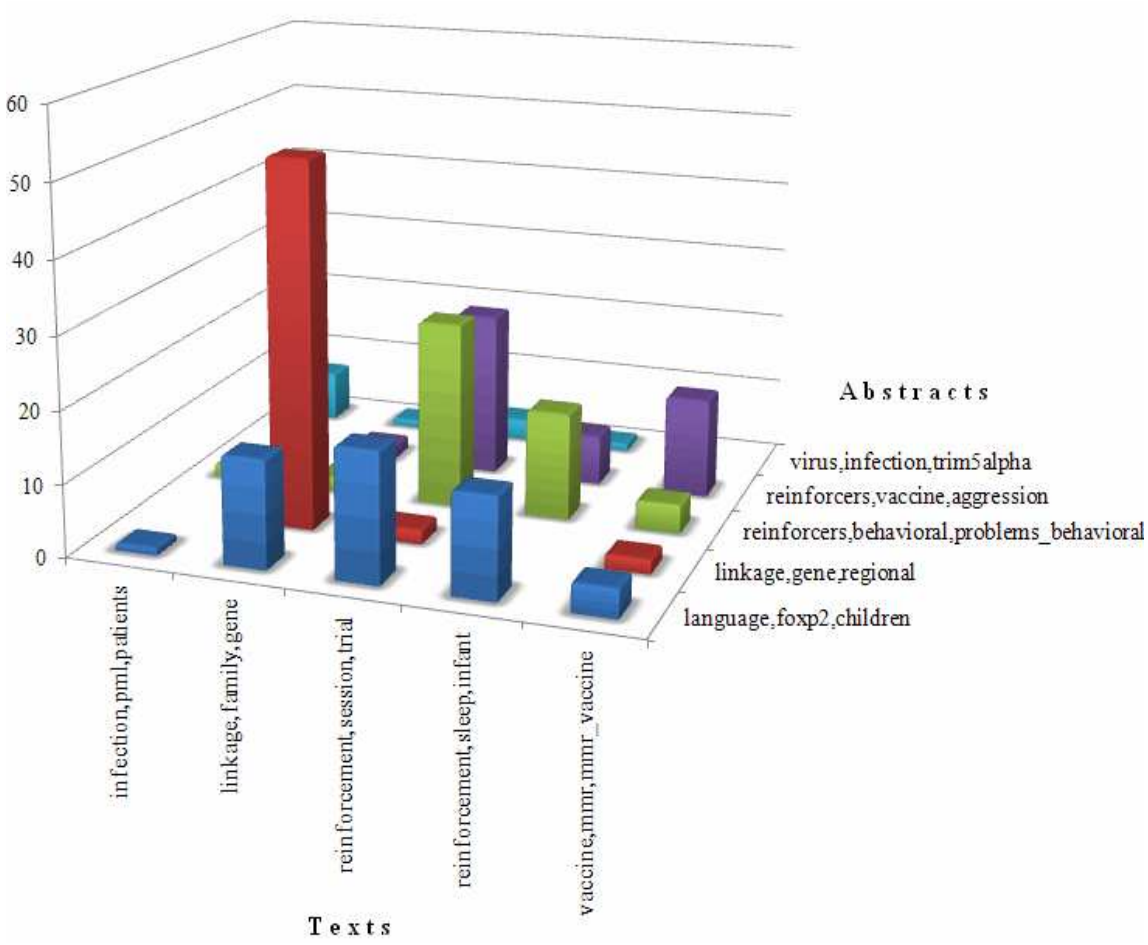


Figure 19: *Comparison between the distributions of documents belonging to the ontology concepts of abstracts and bodies of texts when documents were divided into 5 sub-concepts*.

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 the 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*)[6].

Our graphic presentation of compared ontologies clearly exposes the main clusters of autism articles, which are shown as the highest columns in the graphs in Figures 16 and 17. Besides this, the graphic presentations in Figures 18 and 19 provide a powerful way to visualise the most important similarities between the observed ontologies. One major similarity is identified between the groups of genetic documents, which include the major number of the same articles from the observed dataset. In addition, a relatively large similarity can be seen also between the text and abstract groups that deal with virus infections. Slightly less specific is the similarity between abstracts and texts from the groups: *reinforcement, session, trial* and *reinforcement, sleep, infant*. Although the concept matching presented in Figures 18 and 19 is not completely evident, a general tendency can clearly be found in the diagonal elements. It can be seen that the largest collection of autism documents always deals with genetics. Determining the proper number of top-level concepts (the value of parameter *k*) for a specific domain is

6 To facilitate the recognition and grouping of common words the use of lemmatization or stemming is recommended. Such text preprocessing tasks improve the matching of inflected forms of words based on a common lemma or root.

very important when constructing ontologies in a semi-automatic way. The goal is to find a reasonable compromise between the complexity and comprehensibility of the single-level ontology concepts in the domain. Therefore, experimenting with other values of *k* may also reveal some interesting domain properties.

Our experimental results (Tables 7 and 8) also show that there is a substantial similarity between constructed ontology concepts from abstracts and full texts, while there is less similarity between ontology concepts from titles and abstracts and still less between titles and full texts.

Table 7: *Comparison of individual keywords extracted from concepts names of autism ontologies when documents were divided into 5 sub-concepts.*

Keywords_Titles_5	Keywords_Abstracts_5	Keywords_Texts_5
behavior	behavioral	family
children	children	infant
genetic	gene	gene
autism	infection	infection
linkage	linkage	linkage
reinforcement	reinforcers	reinforcement
children_autism	reinforcers	reinforcement
chromosome	vaccine	vaccine
detection	virus	mmr
disorders	aggression	mmr_vaccine
effects	foxp2	patients
neurodevelopmental	language	pml
social	problems_behavioral	session
spectrum	regional	sleep
syndrome	trim5alpha	trial

Table 8: *Comparison of individual keywords extracted from concepts names of autism ontologies when documents were divided into 8 sub-concepts.*

Keywords_Titles_8	Keywords_Abstracts_8	Keywords_Texts_8
child	children	infant
early	age	prompted
genes	gene	gene
linkage	linkage	linkage
activated	mmr	mmr
analysis	mmr_vaccine	mmr_vaccine
preference	preferred	aggression
assessment	problems_behavioral	problems_behavior
reinforcement	reinforcers	reinforcement
autism	stereotypy	stereotypes
case	stimulus	reinforcement
teach	teach	teaching
vaccination	vaccine	vaccine
children_autism	auditory	chemical
chromosome	behavioral	cortical
disorders	language	executive
effects	mmr	family
evidence	parent	ht
functioning	procedure	legs
schedules	question	nv
social	regional	script
specific	sensory	secretin
susceptibility	sounds	session
syndrome	vaccine	sleep

These findings suggest that titles are not informative enough to be taken as the only source for constructing ontologies. Compared to general knowledge on autism, ontology concepts from abstracts show the highest resemblance. Our results thus confirm the intuitive expectation that constructing ontologies from abstracts is a rational choice when uncovering the structure of a given scientific field. The titles as well as the full texts are typically less useful for the given task. However, when dealing with full texts, some preprocessing tasks such as stemming and stop words removal can improve the utility (Cohen et al., 2005).

The obtained top-level concepts were presented to the expert in autism. As advocated by OntoGen's literature (Fortuna et al., 2006), we renamed the concepts accordingly, based on the suggested keywords. The domain expert found the tables informative and in accordance with her line of reasoning in autism. In particular, the clustering of the selected articles was in most cases fairly intuitive, although the keyword description of some of the generated concepts was not so straightforward. An important confirmation of the resulting ontology construction is also the recent state of autism research as described by Zerhouni (2004), which summarises the main scientific activities of autism research in the major areas of epidemiology, genetics, neurobiology, environmental factors and specific treatments of autism.

Therefore, using tools for semi-automatic ontology construction from scientific articles can significantly speed up the process of becoming acquainted with the domain of interest. Instead of reading an extra load of literature, researchers can first generate top-level domain ontology concepts and thus obtain a general overview and understanding of the domain. After that, a detailed study of the concepts of interest might be in order. In such a way, semi-automatically constructed ontologies actually helped us to review and understand the complex and heterogeneous spectrum of scientific articles about the autism domain.

In addition, our observations from the analyses of the autism domain ontologies supported the filtering of the research results obtained with our literature mining method. Using the MeSH filter in each step of our knowledge discovery process reduces the search results to only show the terms which map to the selected MeSH categories that we choose considering the main concepts identified beforehand in domain ontologies. In fact, in our text mining experiments with the autism literature, we observed that those concepts, which emerge from domain ontologies, enable the establishment of relationships between the key research topics identified in the research domain literature with particular subject categories from the MeSH thesaurus that serve us to filter the results of text mining.

Our graphic presentation of compared ontologies clearly exposed the main clusters of autism articles, which are shown as the highest columns in the graphs. Such thorough examination and comparison of autism domain ontologies revealed distributions of documents inside certain ontology and established relations between titles, abstracts, and bodies of texts. This way we visualized the differences and the similarities in observed ontologies. The major similarities always appeared in collection of those autism documents that deal with genetics. Thus, in the case of a subfield with specific terminology, the experiments showed high similarity between ontologies built on abstracts and ontologies built on bodies of texts.

7 RaJoLink Method for Literature Mining

In this chapter we describe our text mining method called RaJoLink that we developed for discovering implicit relationships hidden in biomedical literature. First, we present an outline of the method focusing on the role of rarity in the open discovery process and on the role of outliers in the closed discovery process, followed by a detailed description of the consecutive steps in the RaJoLink approach. The method is presented decoupled from the application example, which allows for a more general and better structured description. Comments on different choices and their influence on the output of the text mining process are presented as well.

7.1 The role of rarity in the open discovery process

In our approach that we called RaJoLink we have expanded the Swanson's ABC model for literature mining by suggesting how terms *a* can be determined in a semi-automatic way. Although our literature-based discovery approach combines open and closed discovery processes (Figure 20), the main focus is on the open discovery, which is represented in the upper half of Figure 20 with a goal to identify terms *a*.

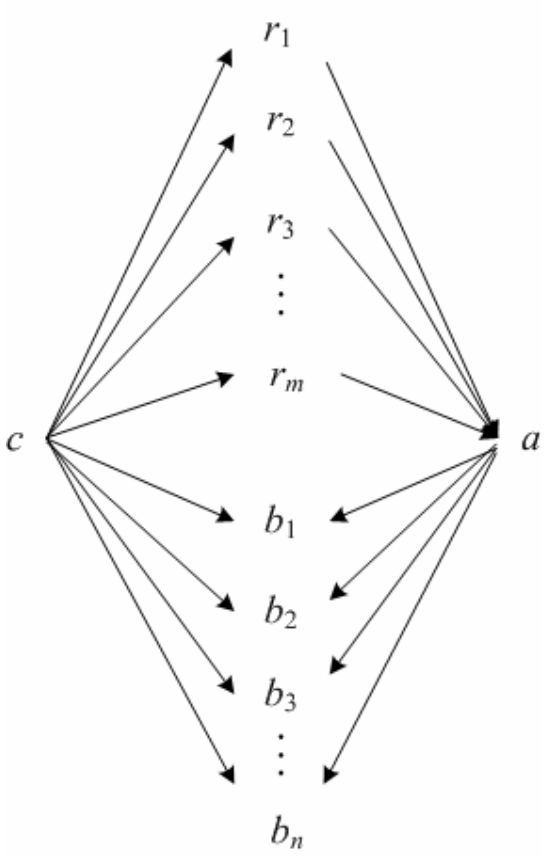


Figure 20: *Combined open and closed discovery process in the RaJoLink method*. The upper half of the figure corresponds to the open discovery (identifying rare terms *r* and finding a joint term *a*) and the lower half to the closed discovery (searching for linking terms *b*).

The main emphasis in text mining investigation has been to directly exploit co-occurrence based relationships between MEDLINE documents (Srinivasan et al., 2004). The open discovery process in RaJoLink is, on the contrary, based on identifying rare terms within literature *C*. Rare terms are our innovative pathways to an unknown term *a*. Each rare term in the RaJoLink method refers to a term that is exceptional for the literature on term *c*. More precisely, a term is rare if it appears in less than or equal to *n* records, *n* being a parameter that can be varied in the experiments. In our case, *n* was set to 1, as the more rare a term is the higher is its possibility to lead to novel connections. The rationale to motivate our choice of *n* is that if a piece of information appears rarely in the set of articles, not many researchers are acquainted with it, so it might be worth exploring it further. Note that the role of rare terms in our approach does not correspond to the role of terms *b* as used in the Swanson's ABC model for relating two disjoint literatures (literature *C* and literature *A*). In our case, rare terms are used to generate term *a* as a joint term that is shared by two or more individual literatures on the selected rare terms.

The reasons for our focus on rare items within literature *C* lay in the associationist creativity theory (Mednick, 1962) with particular regard to the kinds of context-crossing associations, called bisociations (Koestler, 1964). Bisociation involves the literal processes of the mind when making entirely new connections among concepts from contexts or categories of objects that are normally considered separate categories. Throughout the history of science, this mechanism has been the essence of innovative insights and paradigm shifts. However, no comprehensive ICT methodology has yet been developed on this basis. Therefore our aim is to show that the RaJoLink methodology can contribute to this particular approach to scientific discovery, which is based on an existing, but hitherto not computationally implemented notion of bisociation.

Rare observations that appear to be inconsistent with the remainder of datasets are not necessarily characterized by errors but may provide an indication of something unexpectedly useful (Barnett and Lewis 1994). To discover new useful information and interesting knowledge, people have to be constantly involved in a process of creating and evoking latent possibilities, for that reason Magnani (2007) considers humans as chance seekers. Chen (2005) claims that there are even many hard decision-making problems that are solvable by discovery of chance events, which are typically rare.

Rare events may co-occur with important events, while the important events can be obtained by data mining (Wu and Tawfik 2006). As the aim of modern discovery is obtaining the previously unnoticed chances at the intersections of multiple meaningful scenarios, tools for indicating rare events or situations play a significant role in the process of research and discovery (Ohsawa 2006). From this perspective, researchers have to be sensitive to the curious or rare observations of phenomena in order to provide novel possible opportunities for reasoning (Magnani 2005) and be aware of the powerful support that data mining tools can have for choosing meaningful scenarios (Ohsawa 2006). In this regard, Wu and Tawfik (2006) propose an application of combined abductive and analogical reasoning to generate rich knowledge base and to support the discoveries by extension of hypothetical reasoning. Ohsawa, on the other hand, concretely suggests the integration of data mining tools for chance discovery such as KeyGraph together with the Influence Diffusion Model, a method for discovering influential comments, opinion leaders, and interesting terms, which he proved to be successful for discoveries in the hepatitis domain (Ohsawa 2006).

Rarity as a principle has been extensively researched in the field of ecology statistics (Ellison and Agrawal 2005). These investigations include the rarity with which exceptions really occur and they are usually driven by biodiversity and conservation policies (Carney 1997). Considering this, special concern of ecologists has been devoted to studying rare species (Boughton 2001). They recognized two syndromes of rarity: habitat-limited species that were rare because their habitat was rare and dispersal-limited species that were rare because they stayed behind due to a catastrophic turnover of old growth. While ecologists' primary concern was preventing the extinction of rare species, they also identified the potential of dispersal-limited species to adapt to the changed environment.

Historic exception discoveries often evidence the connection between surprise and interestingness (Suzuki and Kodratoff 1998). Rare events actually attract a lot of attention in the research world and are becoming increasingly popular in text mining applications as well. Moreover, exceptions in data often comprise valuable information on abnormal behaviour of the phenomenon described by the data (Aggarwal and Yu 2005).

Detecting meaningful terms that rarely appear in a text collection, can be viewed as searching for the needles in the haystack. This popular phrase illustrates the problem with rarity since identifying useful rare objects is by itself a difficult task (Weiss 2005). In the RaJoLink method the rarity principle is employed in the open discovery process as a means to find new interesting pieces of knowledge that were previously unrelated in the available literature.

Typically, we take that a term is rare if it appears just in one record of the input set of records no matter how many times the term appears within that particular record. The rationale behind it is that if a piece of

information is abundant in the set of articles, it might be speculated that its impact to the field under study is well-covered; however, if it appears rarely, not many researchers are acquainted with it, so it might be worth exploring it further. Besides, Schönhofen and Benczúr disproved the commonly held presumption that rare terms are not informative for text representation (Schönhofen and Benczúr, 2006).

Similarly to dispersal-limited species from ecology, such pieces of information might be either on their way to extinction or might embody a potential for new development in the field. In order to distinguish between the two options, expert guidance is needed in the process.

7.2 The role of outliers in the closed discovery process

In statistics, an outlier is an observation that is numerically distant from the rest of the data, or more formally, it is an observation that lies outside the overall pattern of a distribution (Moore and McCabe, 1999). While in many data sets the outliers are usually regarded as false signals or an imbalance in the data (Lavrač and Gamberger 2001), there are also several examples where the outliers actually led to important discovery of intriguing information. Outlier mining has proved to have important applications in fraud detection and network intrusion detection (Aggarwal and Yu 2005; Lazarevic et al. 2005; Singhal and Jajodia 2006). Similarly, much attention to the study of outliers is paid in the economic field, particularly in finance and business, where rare events can be a sign of interesting unusual activities or observations like, for instance potential sales opportunities (Leung et al. 2006).

A specifically intriguing aspect of outlier detection is emerging within the climate research and extreme weather events prediction. There has been much interest in investigating the impacts, intensity and distribution of rare extreme events over a certain period of time (IPCC 2007). Current attention to rare weather phenomena is driven by their possibility to become regionally more variable or extreme menace to human life, civil infrastructure and natural ecosystems, what may have significant socioeconomic impacts (Frei and Schär 2001).

Examples of outlier relevancy can also be found in social research. Anselin et al. (2007) performed the outlier detection by an explicit spatial analysis of the social indicators such as prenatal care rates, low birth weight, and infant mortality across Virginia counties. With the visualization of extreme values they suggested how outlier analysis could contribute to the discovery of counties with persistently elevated child risk factors over time.

In the closed discovery process of the RaJoLink method, linking terms *b* that bridge literature *A* and literature *C* can be considered as outliers. Having disparate literatures *A* and *C* the RaJoLink method proceeds towards the closed discovery. At that stage, both domains are examined in order to assess whether literatures *A* and *C* can be connected by implicit relations. Relations between *A* and *C* are established via *AB* and *BC* relations. To that end, we automatically retrieve articles on a selected joint term from MEDLINE and consider their analysis together with the analysis of the starting literature *C*. Each document from the two literatures (*A* and *C*) is represented by a set of words using the Bag of Words (BoW) representation (Sebastiani, 2002) and the appearance of co-occurring words is employed as a measure of the content similarity between documents.

The similarity between documents can be determined by calculating the cosine of the angle between two documents represented as BoW vectors (Grobelnik and Mladenić, 2005). This measure is called *cosine similarity* (Grobelnik and Mladenić, 2005) and is commonly used in information retrieval and text mining to determine the semantic closeness of two documents when the document features are represented using a vector space model. This way, all the documents can be listed according to their similarity in the similarity graph (Figure 21).

The visualization of documents in the closed discovery process is supported by using the OntoGen tool (Fortuna et al., 2006). One of its features is its capacity of visualizing the similarity between the selected documents of interest in the document's similarity graph, as illustrated in Figure 21.

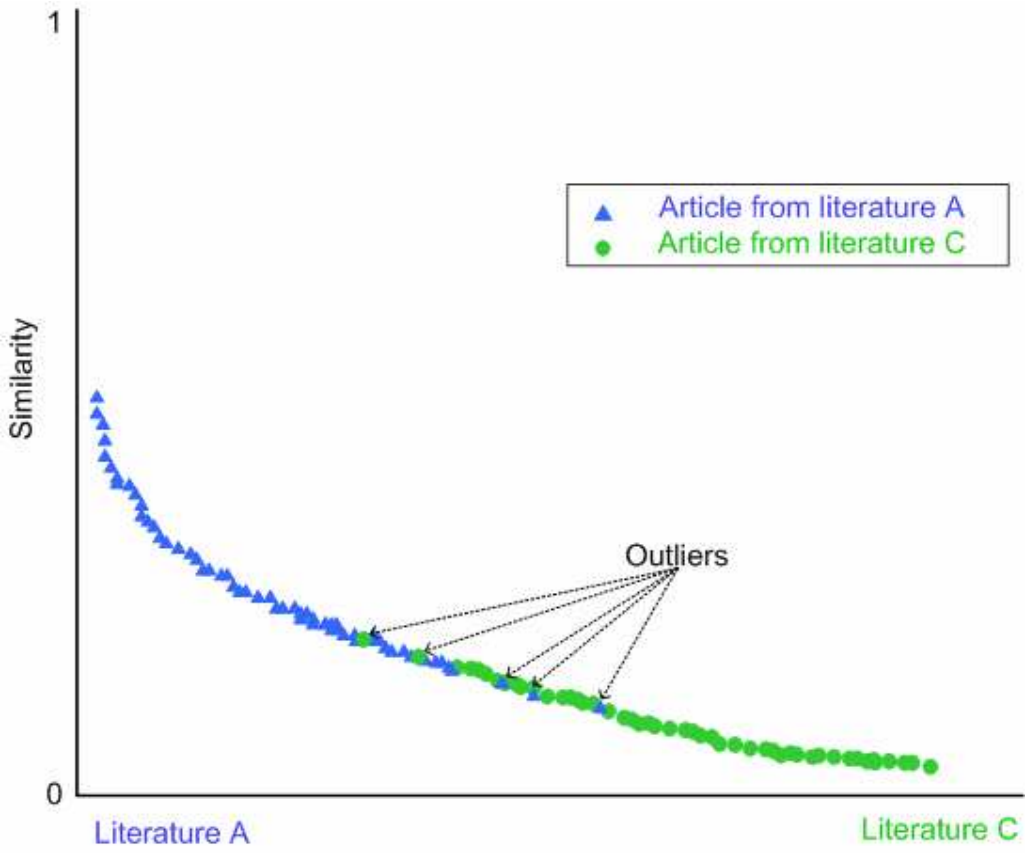


Figure 21: *Similarity graph representing instances of literature A and instances of literature C according to their content similarity*. The distinctive outliers are positioned far enough away from the most typical representatives of the two heretofore unrelated literatures.

The idea of representing instances of literature *A* together with instances of literature *C* in the same similarity graph with the purpose of searching for their links is another unique aspect of our method in comparison to the literature-based discovery investigated by others. By focusing on outlying and their neighbouring documents in the documents' similarity graph that is defined over the combined dataset of literatures *A* and *C*, we perform also the closed discovery phase in an innovative way. In this manner, the outlying documents can be used as a heuristic guidance to speed-up the search for the linking terms and alleviate the burden for the expert.

7.3 Method overview

The entire RaJoLink method involves three principal steps, *Ra*, *Jo* and *Link*, which have been named after the key elements of each step: rare terms, joint terms and linking terms. Therefore, we have called the method RaJoLink after these key procedural elements: *Ra*re, *Jo*int, and *Link*ing terms. The three method's steps are presented in Figure 22. Step *Ra* and step *Jo* together implement the open discovery process, while step *Link* implements the closed discovery.

In step *Ra*, the literature about phenomenon *C* (i.e. the domain under investigation) is examined. The aim of this step is text analysis of the literature about phenomenon *C* in order to identify interesting terms that rarely appear in the documents about phenomenon *C*. In step *Jo*, separate sets of documents about the selected rare terms are inspected and interesting joint terms that appear in the intersection of these sets of documents about rare terms are examined. At least one of them is then selected as the candidate for *A*. The relationship between phenomenon *C* under investigation and a candidate joint term *a* represents the hypothesis generated in the open discovery process of the RaJoLink method. In step *Link*, which implements the closed discovery, linking terms *b* that bridge the gap between the domain *A* and the domain *C* are searched for. Relations between literature *A* and literature *C* are established with pairs of documents

that contain *AB* and *BC* relations. Finally, the user has to evaluate the *AB*, *BC* pairs of documents in support of the generated hypotheses about the relation between the domain *A* and the domain *C*.

In the continuation we present these steps in more detail.

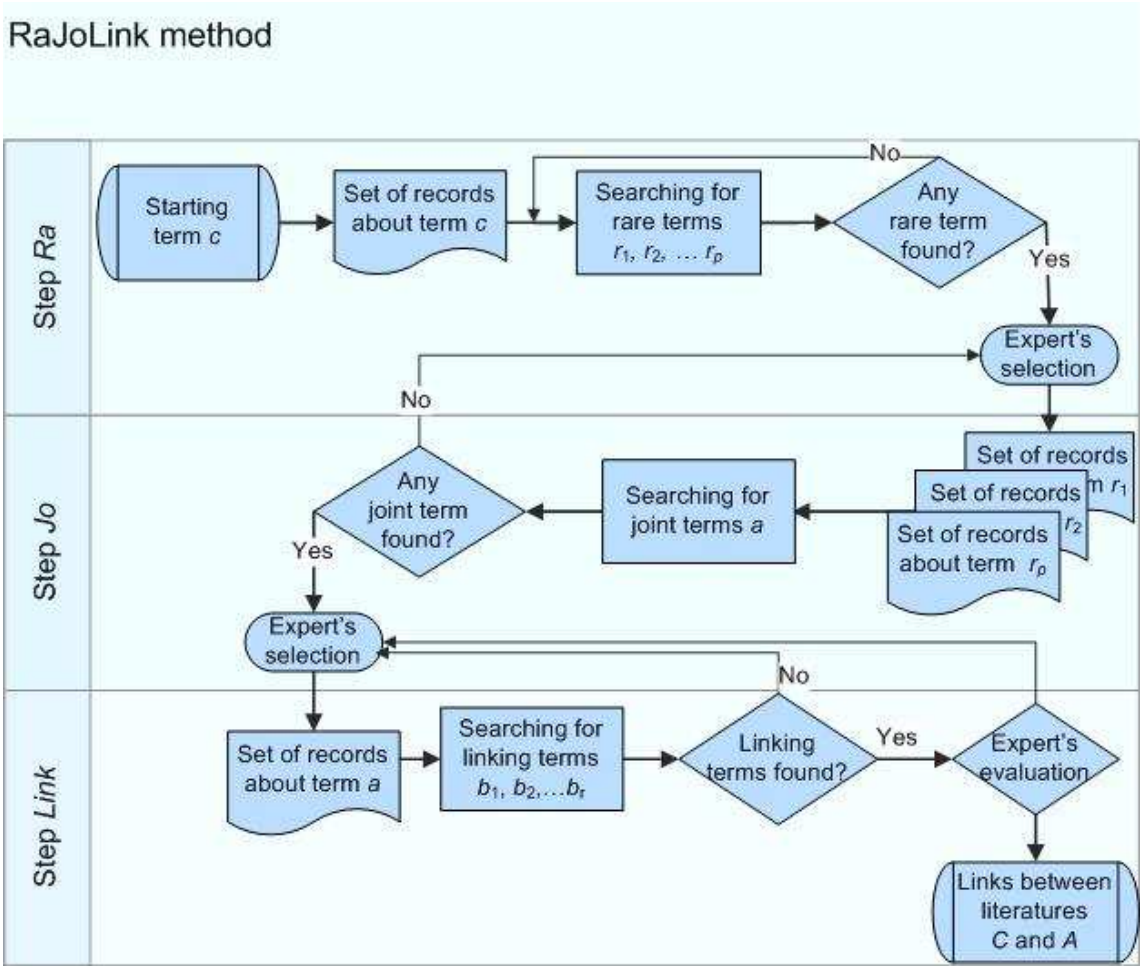


Figure 22: *Flow chart showing the procedures of the RaJoLink method.*

7.4 Step *Ra*

The aim of this step is to identify rare pieces of information in a set of documents about term *c* (the literature *C*) in order to increase the chance of discovering useful implicit relations, which are still unpublished in the literature. Let us denote the total number of records in the input file with *N*, and with *n*(*T*) the number of records that contain term *T*. In practice, the attention is focused on terms that rarely appear in the input set. A term appears rarely in the input set of records if it appears in a relatively small portion of them. Typically, we take that term *T* is rare if *n*(*T*) equals 1. However, note that such constraint is quite sensitive to adding new articles to the input set of records. The term rareness or commonness in text corpus may change by adding new text to the existing input corpus. An infrequent term will become more frequent if the text that was added to the input file contains such a term.

The graph in Figure 23 shows the number of terms with respect to the total frequencies of terms. The total frequency *n*(*T*) of a term *T*, therefore, represents the number of records (e.g., documents) that contain term *T*. While moving along the X-axis, we observe terms from the least frequent to successively more frequent ones. For a given total frequency, the Y-axis value reveals how many terms appear with that given total frequency in the set of records.

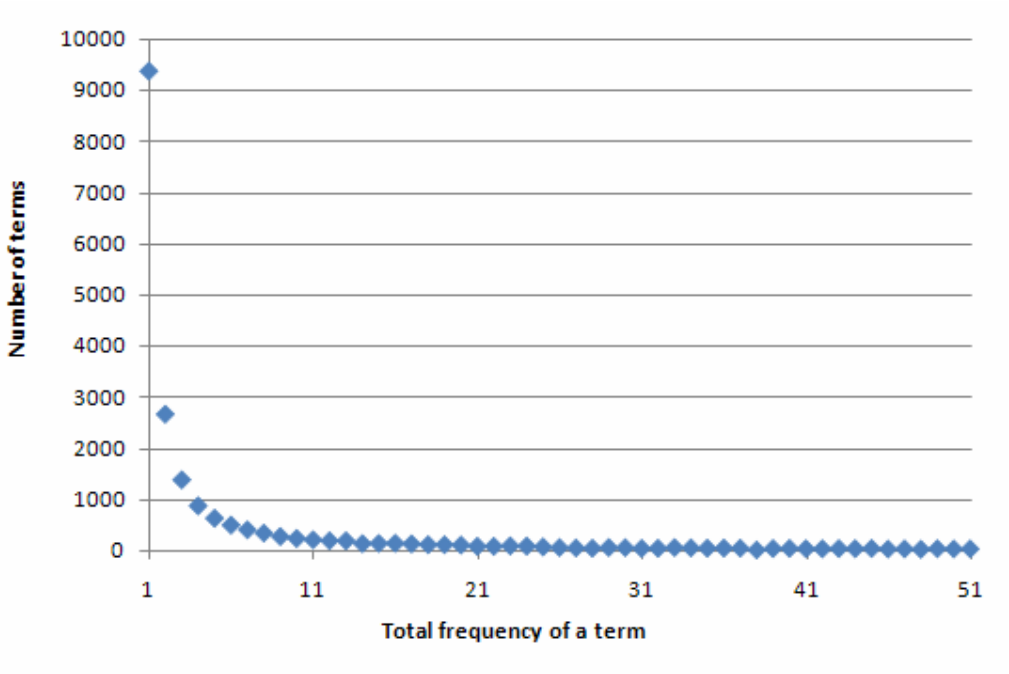


Figure 23: *The number of terms according to their total frequencies in the set of abstracts that were available from 11,781 articles on autism published in MEDLINE until the end of year 2007*. The total frequency of a term (X-axis) represents the number of abstracts that contain that particular term. The Y-axis value reveals how many terms appear with a given total frequency in the set of records.

In the graph in Figure 24 only the terms with total frequency equal to 1 have been considered. These are the terms that occurred in only one record within the set of records about autism. In this graph, their internal frequencies are represented. The internal frequency of a term *T* also known as *term frequency* (Salton and Buckley, 1988) is the number of times term *T* appears within a single record.

In the abstracts of 11,781 articles on autism from MEDLINE, about 7,500 terms have been found that appeared only once in a given record. This means that they not only appeared in just one record from the whole set, but also just once in that record. On the other hand, there have been a small number of terms that occur relatively frequently in a single document. The most frequent of the rarest terms in our experimental case has been the term *GJB2*, which is used for a human gene, also known as gap junction beta-2 or connexin 26 (Wiley et al., 2006). It appears 18 times in a single abstract on autism and exclusively in that one article on autism.

While the graphs in Figures 23 and 24 show different statistics the distributions reveal substantial similarity. The most terms have both frequencies equal to 1, which is not surprising. They are natural candidates for rare terms. However, in order to make the process more efficient, additional mechanisms like filtering according to Medical Subject Heading (MeSH) should be applied.

To implement filtering of terms there are alternative options that can be chosen by the user, which can be either one or more top-level and/or second-level categories from the MeSH hierarchy. Words that are of high interest in the medical science are included in Medical Subject Headings. Focusing on such words can narrow down the search space and, thus, speed-up and improve the inference process.

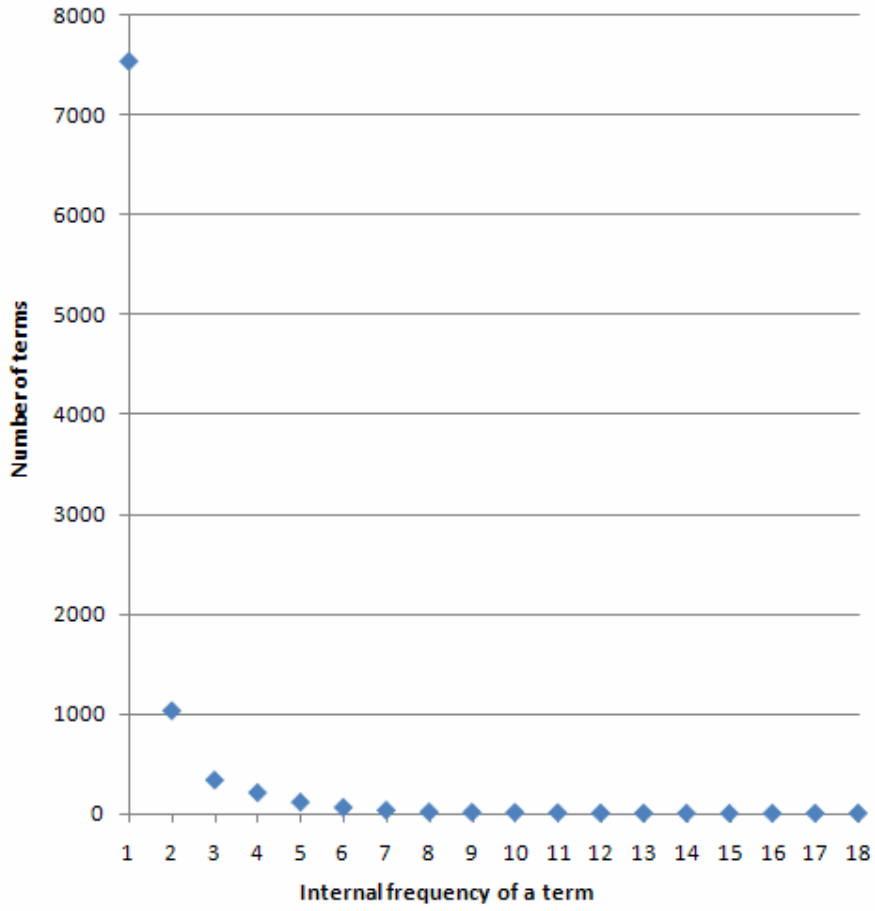


Figure 24: *The number of rare terms according to their internal frequencies observed in the set of abstracts that were available from 11,781 articles on autism published in MEDLINE until the end of year 2007.* The internal frequency of a term (X-axis) is determined by the number of times the term appears within a single abstract.

In order to make the search for rare terms effective, the preprocessing step includes lemmatization, exclusion of words from a stoplist and filtering according to MeSH classification. The presented method first compares three-word terms from the input text with the 2008 MeSH terms. If a multiple word term is found among MeSH terms, it is added to the list of terms for the further statistical analysis of strings. Afterwards, the same is executed with two-word terms from the input text. Multiple word terms that are not found in MeSH thesaurus are treated as individual words. We use the second-level categories from the 2008 MeSH tree structure (e.g., Behaviour and Behaviour Mechanisms- F01, Psychological Phenomena and Processes - F02, Mental Disorders - F03, Behavioural Disciplines and Activities - F04) to classify terms from the input text collection. Each of the second-level categories belongs to one of the top-level categories in the MeSH hierarchy.

On the other hand, words that are not part of the MeSH, are automatically added to the second-level general category, which we named Various - V05 and added to the second-level categories V01, V02, V03, and V04 of the top-level MeSH category: Publication Characteristics - V. For this reason we use the top-level category also named Various (Figure 25) instead of the originally named top-level MeSH category (i.e. Publication Characteristics).

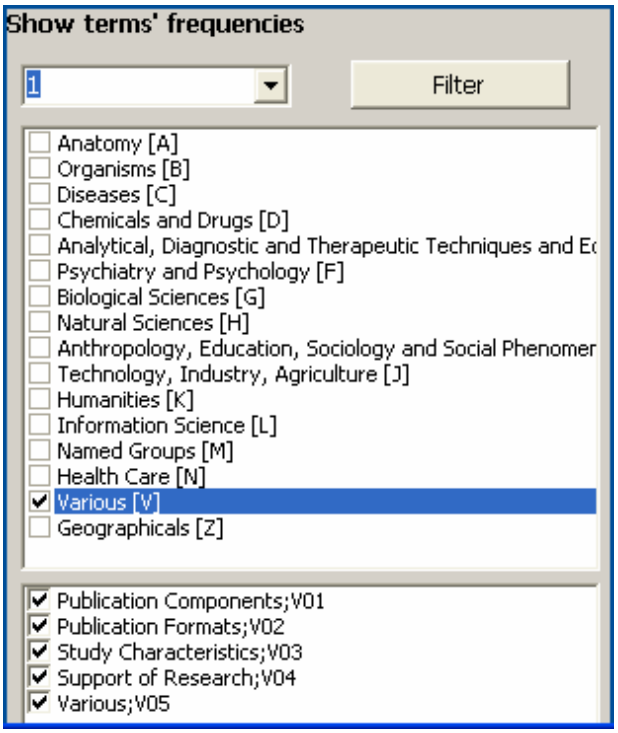


Figure 25: *A screenshot of the RaJoLink system showing second level MeSH categories (V01, V02, V03, VO4) and the general category of terms (V05) within their top-level category (V).*

Lemmatization is the morphological analysis of words that is used to correctly identify the lemma for each word and in this way eliminate various forms of a single word. We employed the lemmatizing software from the LemmaGen library developed by Juršič and colleagues (Juršič et al., 2007).

Stoplists contain words that are predictably of no interest and should, therefore, be excluded from the input records. In this way, all the terms that are not subject-oriented should be ignored. We use a list of 571 English stop words (i.e. generic words such as a, able, about, above, and according).

For each term *T* that appears in the set of records, *n*(*T*) is calculated using frequency statistics based on Bag of Words (BoW) text representation (Sebastiani, 2002), wherefore we employed the Txt2Bow utility from the TextGarden library (Grobelnik and Mladenić, 2004). As previously stated, all terms with *n*(*T*) equal to 1 are selected as rare. Let us denote such terms with *r*. After that, the domain expert has to indicate interesting rare terms $r_1, r_2,\ldots r_p$ out of them, with regard to the background knowledge about the subject of interest. Note that *p* can be regarded as a parameter of the method.

When rare terms cannot be found by using this minimum frequency of term's occurrence, it is necessary to return back to the beginning of this step and repeat the process by taking into account larger number of input documents or by choosing higher value of parameter *n*(*T*). The appearance of the selected rare terms in the literature *C*, however, means that they had already been reported in the context related to term *c*. Therefore, the rare terms represent only intermediate results towards the new knowledge discovery. Neither should they be considered to have the same property as *b* terms, although the rare terms individually co-occur with the literature *A* and the literature *C*, but not with *A* and *C* jointly. The reason is that when testing the hypotheses by searching for meaningful linking terms *b* between the literatures *A* and *C*, the focus should be on the most frequent terms, while the rarest ones are interesting in the hypotheses generation phase.

7.5 Step *Jo*

When the data analyses described in step *Ra* are completed, new individual sets of records, one for each selected rare term, have to be obtained and further analyzed. For this purpose, a set of articles about each of

the rare terms selected in step *Ra* are automatically retrieved from MEDLINE or extracted from other document source. After preprocessing each set of articles as in step *Ra*, the goal is to find the terms that the text collections have in common. Again, BoW representation is used for the task. For taking into account multi-word terms, the maximal length of n-grams being 3 can be used as standard set of parameters for Txt2Bow utility (Grobelnik and Mladenić, 2004). A term from the BoW qualifies as a joint term if it appears in at least two sets of records about the rare terms. At the same time, it has to be absent from the set of records about term *c*, generated in step *Ra*. The intermediate output of this step is, therefore, joint terms $a_1, a_2,\ldots,a_q$ as the intersection of the literatures on the rare terms. If joint terms are not obtained via the rare terms selected in step *Ra*, it is necessary to return back to the results of the previous step and broaden or change the actual selection of rare terms and repeat the process.

The expert role is crucial also in step *Jo*. Based on the expert's opinion, one of the proposed joint terms is selected for further investigation to be done in step *Link*. On the basis of the selected joint term new hypotheses are formulated and subsequently tested following the Swanson's ABC model for closed discovery.

7.6 Step *Link*

In order to provide explanation for hypotheses generated in step *Jo*, our method searches for links between the literature on joint term *a* and the literature on term *c*. This step is equivalent to Swanson's closed discovery (Swanson, 1990). Nevertheless, our closed discovery approach contains another unique aspect of our method in comparison to the literature-based discovery investigated by others. It is the focusing on neighbouring documents in the documents' similarity graph (Figure 26), which is defined over the combined dataset consisting of literatures *A* and *C* as we extensively analyzed in Section 7.2.

Within the whole corpus of the text dataset consisting of literatures *A* and *C*, which acts as input for step *Link*, each text document represents a single record. After preprocessing similar to that used in steps *Ra* and *Jo*, a single document is represented by a set of words using the BoW representation. The appearance of co-occurring words is employed as a measure of content similarity. Its computation is performed with OntoGen, which was designed by Fortuna and colleagues for interactive data-driven construction of topic ontologies (Fortuna et al., 2006).

The content similarity is based on the textual description of documents and is measured using the standard TF*IDF (term frequency inverse document frequency) weighting method (Salton and Buckley, 1988). This way, all the records are sorted according to similarity and the content related documents are obtained by comparing neighbouring documents from the list (Figure 26). Since the search for the linking terms can be combinatorially complex (Swanson, 1990; Hearst, 1999), focusing on neighbouring documents can be used as a heuristic guidance to speed-up the process and alleviate the burden on the expert to sort out the meaningful explanations.

Some articles in MEDLINE are represented with titles, some with titles and abstracts, but only few of them have full text content available. Any of these articles' parts can be used as an input for RaJoLink. When the input for RaJoLink (e.g., titles, abstracts or full texts) is being prepared, the choice of document parts that are taken into account has an important impact on the guidance of the entire text mining process. In this step, using abstracts rather than entire texts of articles is recommended. This recommendation can be argued by one of our previous studies (Petrič et al., 2006a), which pointed out that using only abstracts produced better results than using whole texts or only titles. Besides, the terminology used in abstracts has a stronger importance when compared to a randomly chosen part of the article's text corpus of the same size. In our supposition, the linking terms $b_1, b_2,\ldots b_r$ that are searched for by word intersections also have an even stronger connecting role between the two literatures this way.

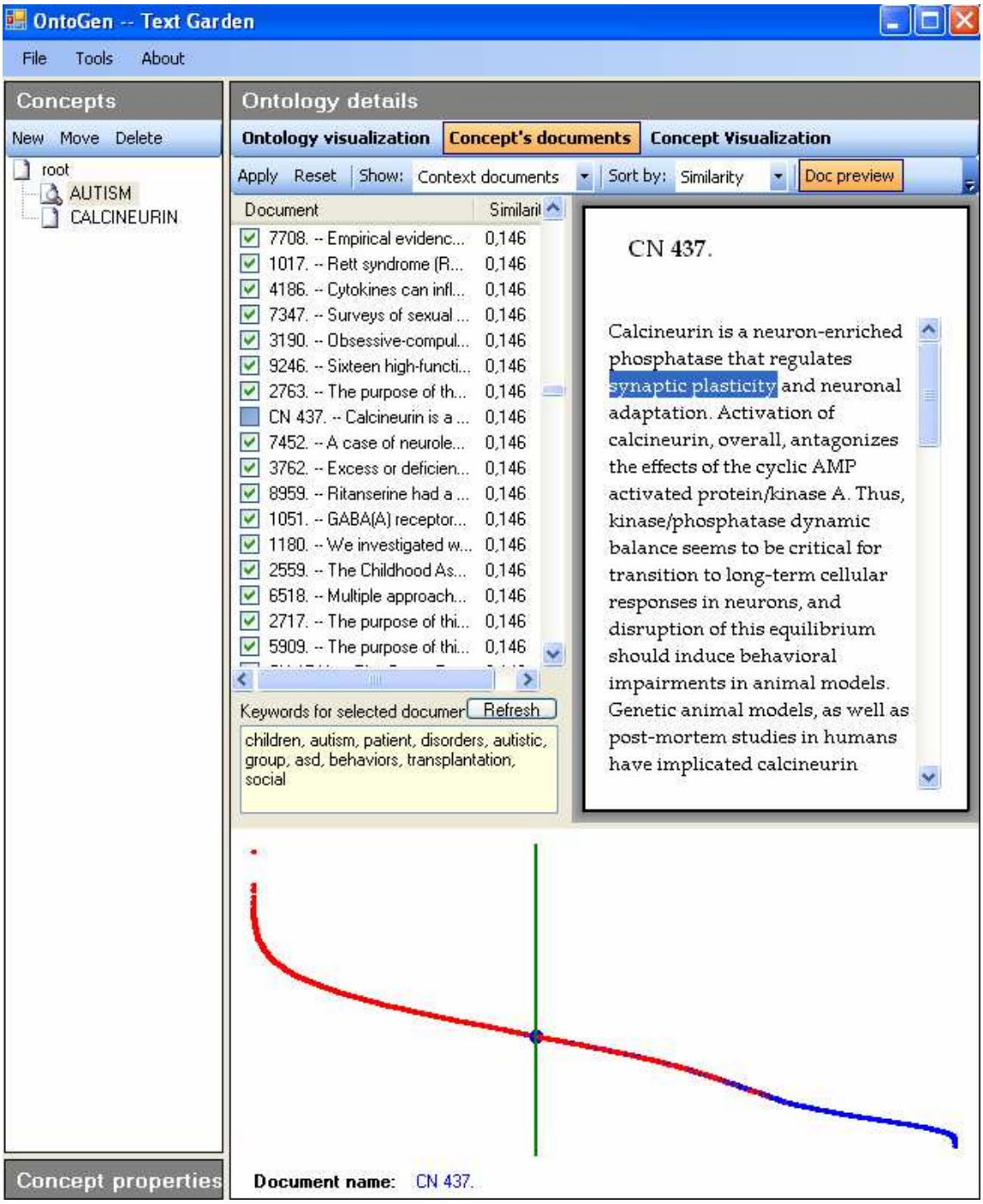


Figure 26: *OntoGen's similarity graph of a set of autism and calcineurin articles' abstracts*. Two main article topics (*AUTISM* and *CALCINEURIN*) are listed on the left side of the window. As the autism topic is selected, the list of abstracts, which are in the relationship with this selected topic, is presented in the central part of the OntoGen's window. The distinctive calcineurin article (*CN 437*) is visualized among the autism context documents.

The presented linking approach suggests the novel way to improve the evidence gathering phase when analyzing individual *a* terms in their potential connection with term *c*. In reality, even Srinivasan and colleagues, who declared to have developed the algorithms that require the least amount of manual work in comparison with other studies (Srinivasan et al., 2004), still need significant time and human effort for collecting evidence relevant to the hypothesized connections. In the comparable RaJoLink's approach, the

domain expert should be involved only in the conclusive actions of step *Link* to accelerate the choice of significant linking terms. The presented procedural elements of steps *Ra*, *Jo*, and *Link* are recapitulated in the schema of Figure 27.

Step	Input	Action	Tool, technique	Expert's involvement	Output
Ra	Set of records about *c*	1.1 Extraction of texts	Digital document archives		
		1.2 Data collection preprocessing	Word processing software		
		1.3 Identification of rare terms *r*	Word frequency statistics		
		1.4 Terms filtering	Content filtering	Indication of interesting rare terms	Rare terms r_1, r_2,…r_p
Jo	Sets of records about r_1, r_2,…r_p	2.1 Extraction of texts	Digital document archives		
		2.2 Data collections preprocessing	Word processing software		
		2.3 Search for joint terms	Word frequency statistics	Selection of a significant joint term	Joint term *a*
Link	Joint set of records about *a* and records about *c*	3.1 Extraction of texts	Digital document archives		
		3.2 Data collection preprocessing	Word processing software		
		3.3 Identification of content related *A* and *C* records	Text analysis		
		3.4 Search for linking terms *b*	Word intersection	Selection of meaningful linking terms	Linking terms b_1, b_2,…b_r

Figure 27: *Schema of the RaJoLink method.*

8 RaJoLink System

In this chapter we present the RaJoLink system. The RaJoLink system has been designed to support the extraction of information from scientific articles and to advance automated processing of such information in order to provide the basis for connecting disjoint literatures. The RaJoLink system provides a framework for literature-based discovery from texts written in English. The system has been developed with Borland Delphi and uses two knowledge sources for its applications in biomedical domains: Medical Literature Analysis and Retrieval System Online (MEDLINE®), and Medical Subject Headings (MeSH®).

8.1 The RaJoLink system description

In order to guide the search towards results interesting for the user, the system is interactive and gives to the user the possibility of making his/her own selection out of suggested terms at each of the steps. One of the important features of the system is also the ability to show to the user a list of suggestions along with the information he or she might find useful in making the selection.

The literature-based discovery in RaJoLink typically starts with step *Ra*. The aim of this step is to find a list of rare terms to support the open discovery process. For this purpose, the frequency statistics based on Bag of Words (BoW) text representation is used, as described within this section. The system searches the MEDLINE database to retrieve literature on the starting term *c*. The search for term *c* can be specified as a single word or a phrase containing several words. In the later case, search terms are combined using logical operators AND, OR, NOT with capital letters. When the user specifies the starting term (word or phrase) and the parts of texts that should be considered (titles or abstracts), the system searches MEDLINE and automatically retrieves the resulting set of records. The user has also the possibility to retrieve texts from a specific file rather than from the MEDLINE.

After obtaining the set of records about the starting term *c*, the preprocessing phase takes place. Preprocessing includes deleting graphics, paragraph marks, and manual line breaks from the texts, so that each document occupies exactly one record in the input file. To effectively build a list of terms from a set of records about the starting term *c*, the preprocessing step includes also lemmatization and exclusion of words from a stoplist.

Lemmatization is used to eliminate inflected forms of a single word. Stoplists contain words that are predictably of no interest and should, therefore, be excluded from the input texts. Hence, the terms that are not subject-oriented should be ignored. After preprocessing phase, the text corpus is transformed in a manner that each record is represented by the set of its terms and terms' frequencies. Following, the total frequencies of terms in the whole text corpus are computed.

To the output list of terms, containing terms and their total frequencies $n(T)$, the Medical Subject Headings (MeSH) codes have been added as qualifiers. Consequently, the filtering according to the MeSH classification can be performed in order to restrict further investigation to only those medical subjects that are of specific interest. The sample output is provided in Figure 28 showing the results of step *Ra* where only rare terms from autism documents were displayed and additionally filtered by MeSH codes. In fact the terms displayed are those with their total frequency equal to 1 and appertaining to the restricted Medical Subject Headings.

Regarding the total frequency $n(T)$ statistics, the rare terms are identified among the terms extracted from sets of records about the starting term, *c*. Principally, the attention is focused on terms that rarely appear in the input set of records. Since the number of such terms can be very big, additional MeSH filtering is used to substantially reduce the potential candidates.

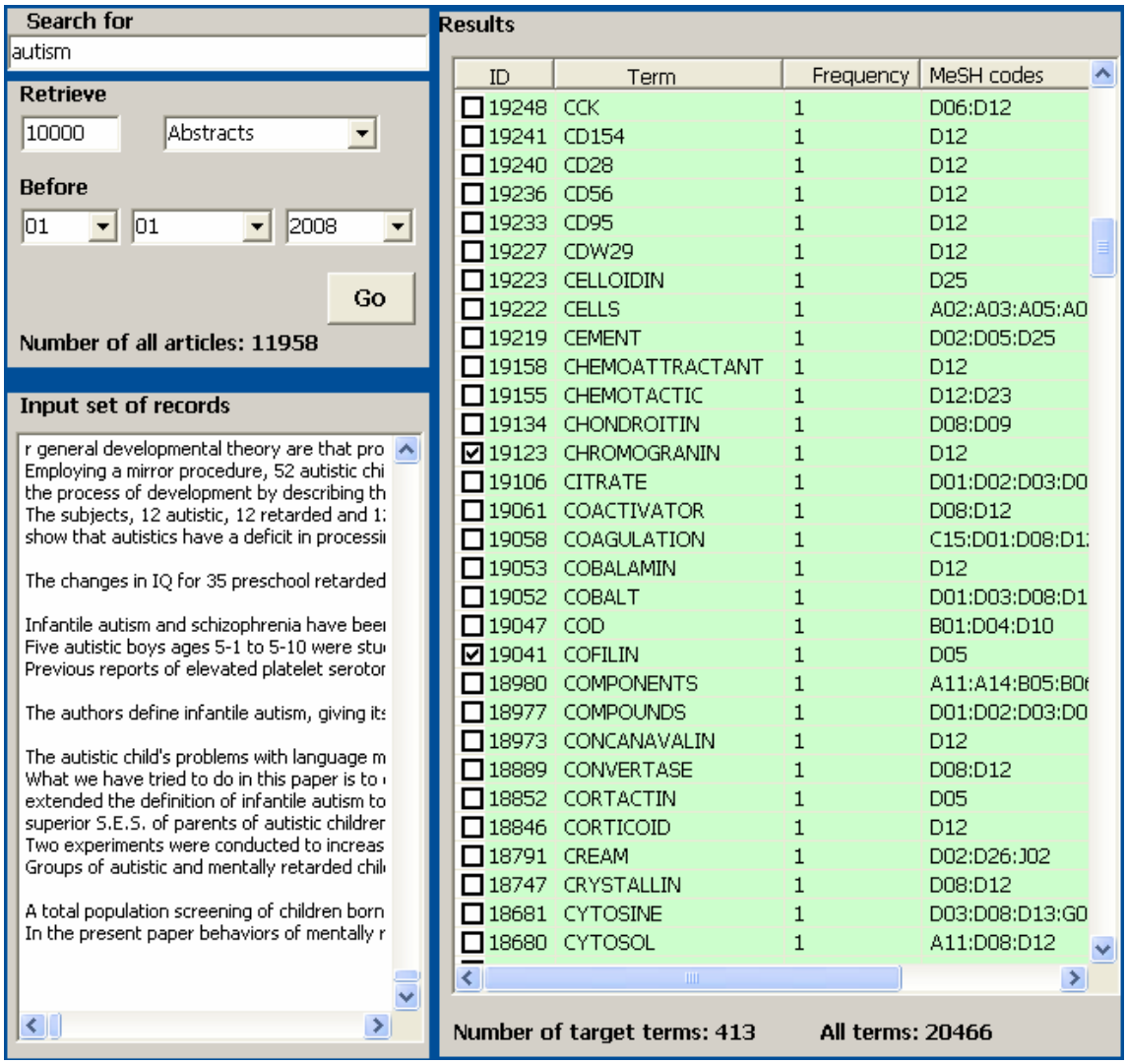


Figure 28: *A screenshot of RaJoLink where only rare terms from autism documents are displayed (parameter Frequency =1)*. MeSH codes for each term are listed on the right hand side. The terms chromogranin and cofilin are selected (checked) for further analysis.

The domain expert defines a subset of rare terms relevant for the further investigation in the following step *Jo*. The computing applied in step *Jo* is an extension of the algorithm for executing step *Ra*. As a consequence, the system looks up for sets of records about each of the rare terms, r_1, r_2,…r_p, which have been selected by a subject expert. To retrieve literature on each of the rare terms, r_1, r_2,…r_p, it again searches in the MEDLINE database.

This step involves also combining frequency statistics for literatures on rare terms and literature on the starting term *c*. Based on the Bag of Words text representation, total frequencies $n(T)$ of terms are computed for each set of records about the selected rare terms. This way, the numbers of records (e.g., documents) in the whole corpora, which contain a term *T*, are counted. The list of terms is therefore composed of terms, their total frequencies in sets of records for each of the selected rare terms and their total frequency in the set of records about the starting term *c*. This last frequency is automatically added to the list of joint terms in order to facilitate the validation of terms regarding their possibility to be selected as a promising joint term, *a*. Actually, the term is valid as a potential joint term only if it has not been reported yet in the literature on the starting term *c*. Therefore the last frequency, which is the total frequency $n(T)$ of a term *T* in a set of records about the starting term *c*, should equal 0. This means that none of the records

from the whole corpus of records about the starting term *c* contains a term *T*. Thus, only those terms that have their last frequency equal to 0, pass the criterion of possible hypothesis generation. However, the higher number of occurrences of a joint term *T* in the sets of records about the selected rare terms means the better matching of a term *T* as the desired joint term *a*. On the other hand, any joint terms that have their total frequency in the set of records about the starting term *c*, equal to 1 or more, should be judged as unlikely to provide worthwhile knowledge discovery in the problem domain.

An example of searching for a joint term is illustrated in Figure 29. It lists the numbers of records in the abstracts of articles on calcium channels, chromogranin, cofilin, lactoylglutathione, and on synaptophysn, which contain the listed terms (e.g., ethylmaleimide, electophoretic, and others). For instance, the term calcineurin appears in 5 sets of records; in 18 articles on calcium channels, in 6 articles on chromogranin, in 8 articles on cofilin, in 1 article on lactoylglutathione, in 4 articles on synaptophysn, but in none of the articles on autism.

In fact, the system searches for the potential joint term also in the set of articles on the starting term (term *c*), which was autism in this example, and calculates the total frequency of the term in the starting set of articles. In the case we took under consideration the word carbon, which is actually not shown in the screenshot in Figure 29, we would see that the total frequency of the term carbon in the set of abstracts of 10,000 articles on autism equals 4. This means that the term carbon appears in 4 records about autism. Consequently, as it has been already mentioned in the literature on autism, it is not interesting for new discoveries in the autism domain. Therefore, we rather search for terms that have not been reported in the literature on autism yet. In the example list of terms, such a term is calcineurin, which has the frequency of the term in the set of articles on autism equal to 0.

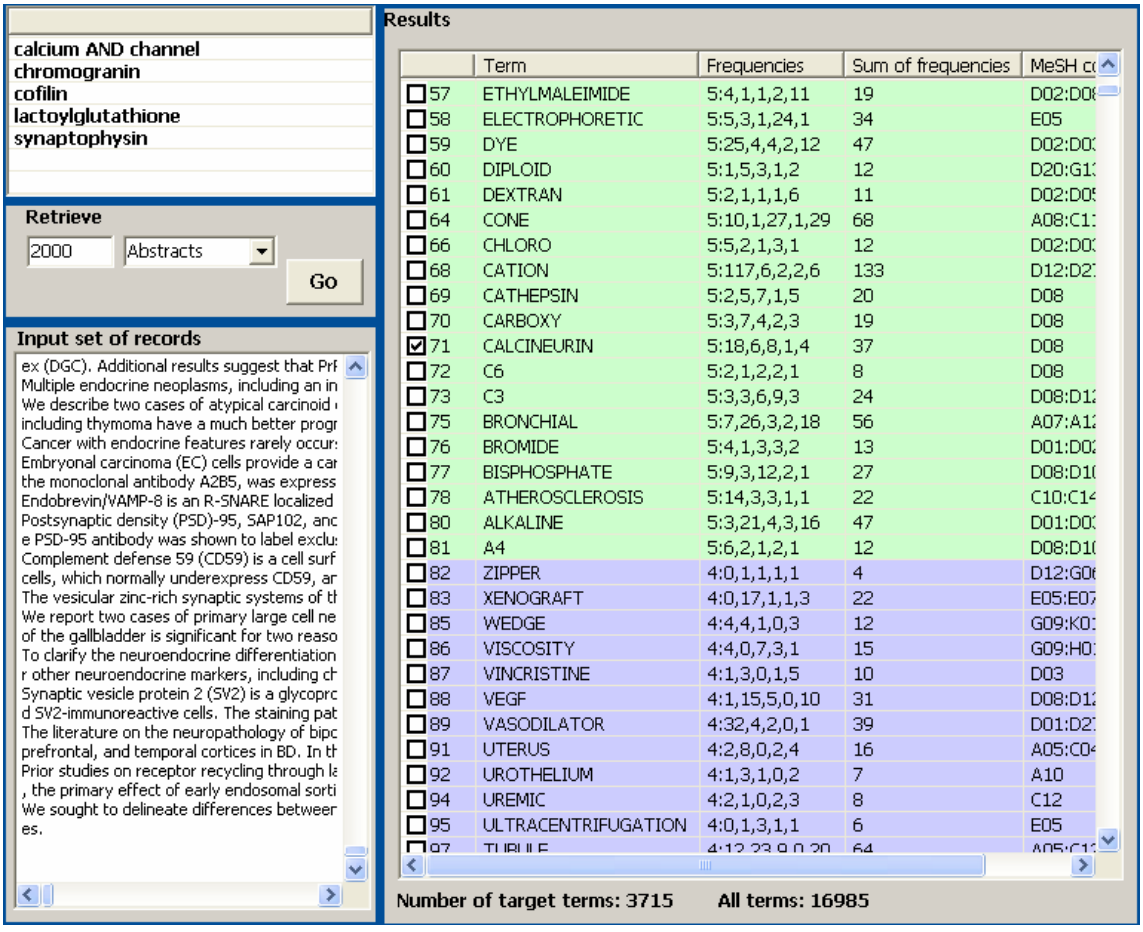


Figure 29: *A screenshot of RaJoLink showing part of results for candidate joint terms*. The term calcineurin is selected for further analysis.

The obtained list of candidate joint terms with their frequencies is arbitrary, since each term can be taken as relevant for the generation of new hypotheses only if the last of all total frequencies equals to 0. As a result, only the candidate joint terms that have their last total frequency value, which is computed for the appearance of a given candidate term within the set of articles on the starting term (term *c*), equal to 0, should be considered for further analysis. The selected joint term *a* is then considered in the last step (*Link*) for the detection of implicit links with the problem domain denoted by term *c*. In this manner, a substance calcineurin has been chosen as a joint term in our experimental case of autism (Petrič et al., 2007; Urbančič et al., 2007).

Having disparate literatures *A* and *C* the user can proceed towards the closed discovery. At this stage, both domains are examined in order to assess whether the literatures *A* and *C* can be connected by implicit relations. To this end, the system retrieves articles on a selected joint term *a* from MEDLINE. Similar to the Swanson's closed discovery approach, the search for linking terms consists of looking for terms *b* that can be found in both sets of records, i.e., in the records on term *a* as well as in the records on term *c*. Finally, pairs of records (*A*,*C*) containing the same term b have to be inspected, to confirm the relevancy of terms *b*. Therefore, a subject expert should assess whether the jointly considered statements in given pairs of records can really support the hypothesis about novel relation between *a* and *c*.

8.2 Future development of the RaJoLink system

There are two main complex areas in which the RaJoLink system requires some further work. The first is about matching ambiguous, idiosyncratic terms. In medicine, this is often the case of acronyms, abbreviations, symbols and synonyms for chemical names. Therefore, other text analysis methods should be comprised, for automated identification of semantic variants such as abbreviations, acronyms and synonyms. We plan to use contextual features and external resources (e.g., thesauri and ontologies) that might resolve such ambiguities for RaJoLink. In particular, WordNet, a lexical database for the English language (Miller, 1995) can be used to disambiguate semantic variants of terms before calculating frequencies of semantically related terms in the text collections.

Secondly, some further work also remains in evaluating the results at step *Ra* of our method, which is necessary for proper prediction of relevant terms for further analysis at steps *Jo* and *Link*. The actual RaJoLink system provides filtering according to MeSH that enables the user to limit search results to one particular or several medical subjects and to the selected maximal number of terms frequency. However, although such filtering is extremely useful for narrowing down the search space it would be necessary to apply additional measures for predicting the quality of results obtained by the RaJoLink method (e.g., against some standards) and this way improve the inference capabilities of the method. This issue could be resolved by the development of approaches that would allow incorporation of pre-existing text mining standards developed on independent domains. Gold-standard approaches of this kind would construct a benchmark for validation of the literature-based text mining results. This way a common evaluation framework would be provided, which could quantify the results of the literature-based text mining applied to a general biomedical domain. Comparing the results obtained in a real experiment to the results of the general evaluation framework would allow for refinement of the experimental design or would highlight interesting areas for further investigations.

For further development of the RaJoLink system we also want to provide more guidance to the users. In future work we will implement different visualizations of results in the RaJoLink system. This way, the system may become an even more efficient knowledge discovery tool in biomedicine.

An example of an advanced solution to this challenge is illustrated for step *Link* in Figure 30. We suggest that results should be visualized as pairs of the articles' parts retrieved from MEDLINE that would link specific findings from the literature *C* to the hypothetically related observations in the literature *A*. The novel hypothetical relations between *A* and *C* would be represented with highlighted linking terms in order to provide the basis for quick expert's choice of valuable results.

An additional examination should further investigate the role of parameters' values such as the threshold frequency of terms to be marked as rare, the number of selected rare terms, etc. Moreover, as the RaJoLink system is generally applicable, we will test it also on other problem domains.

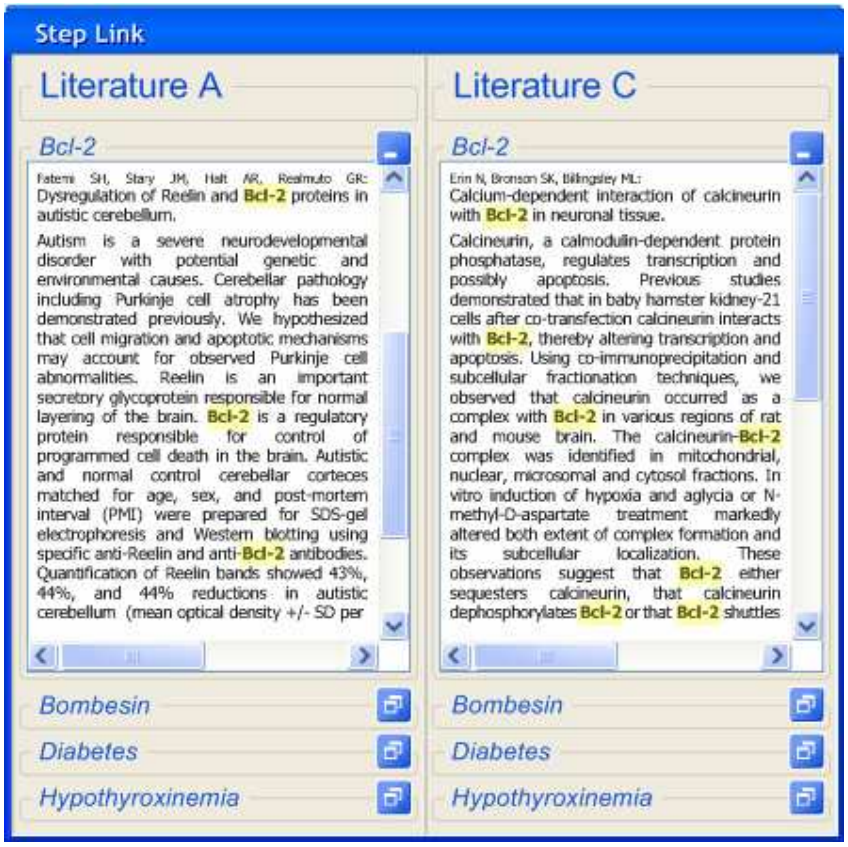


Figure 30: *A screenshot of the proposed user interface showing the visualization of results for candidate linking terms.* Each candidate linking term would be highlighted in pairs of articles from literature *A* and from literature *C*.

Nevertheless, the RaJoLink system already implements the RaJoLink method in an easy-to-use way, appropriate also for general usage. The proposed approach is illustrated in the continuation by its application in the autism domain, where we managed to generate new medical hypotheses from MEDLINE articles with the automated support of the RaJoLink system.

9 Application of the RaJoLink Method to the Literature on Autism

This chapter is dedicated to a practical application of the RaJoLink method to the text analysis of biomedical scientific documents. We present how text mining and link analysis techniques, which are implemented in our method, can be performed utilizing the infrastructure developed in this thesis and show how they can be applied to a biomedical domain. For the experimental field we chose autism, for which causes and risk factors are still poorly recognized although it is known, that both genetic and environmental factors influence this disorder.

9.1 Experimental setup

Autism is a complex, heterogeneous domain and object of investigation in several subfields, which use different procedures for its examination: observations of behaviours and responses to visual, auditory, social and other stimuli, treatment of genetic factors, magnetic resonance imaging, questionnaires, and others. The question of how to connect partial results of individual sciences into a complete picture for now remains an unsolved challenge.

To examine the autism domain we used the scientific literature published on the subject of autism that can be found in the MEDLINE bibliographic database. We retrieved the documents for application of the RaJoLink method to the autism literature using the PubMed search engine. We performed a PubMed search of articles on autism for the period 1998–2007 and thus collected 214 documents with their full text available in the PubMed Central.

The reason for taking full texts of articles instead of abstracts or titles is that the rare terms, which are encountered in articles' abstracts, might be mentioned also in some full texts of articles but not in their abstracts due to the low importance for the subject of the article. In fact, in such cases these terms would not all be the rarest ones for the domain under study. On the other hand, many of the truly rare terms would be overlooked if we observed only abstracts or titles instead of full texts.

After analyzing the literature on autism to gain background knowledge about the autism domain, we generated its statistics of terms and a compiled list of rare terms. With the RaJoLink system, about 2000 terms were fully automatically detected in step *Ra*. Each of them appeared only in one of the documents from our set of 214 autism documents. This means that only terms with $n(T)$ equal to 1 were selected as rare, no matter if they appeared once or several times in that particular document. However, according to Schönhofen and Benczúr a feature instance is rare if it is present in up to 10 documents (Schönhofen and Benczúr, 2006). Therefore, the parameter $n(T)$ could be set to a higher number of documents in case there are no meaningful, interesting results when $n(T)$ is equal to 1.

We disregarded the terms that are not subject-oriented, such as the words: *atomic*, *bundle*, *checkout*, and *dipper*. In our case study, there were more than 9,000 terms, without stopwords (that are automatically deleted), which each appeared only in one abstract from the whole corpus. These account for nearly 45 % of the whole list of different terms that were found in the set of records about autism. As such a long list of terms can still be time consuming and confusing for further analysis, we execute filtering according to MeSH classification by choosing one or more top level or second level MeSH categories. For instance, suppose that our main interest was in enzymes and coenzymes that could influence the phenomenon under research. In this case we would choose only the MeSH category D08, to which enzymes and coenzymes are designated as illustrated also in Figure 31. By such filtering the number of rare terms in our experimental case decreases from more than 9,000 of all rare terms to 264 rare terms that refer to organic chemicals.

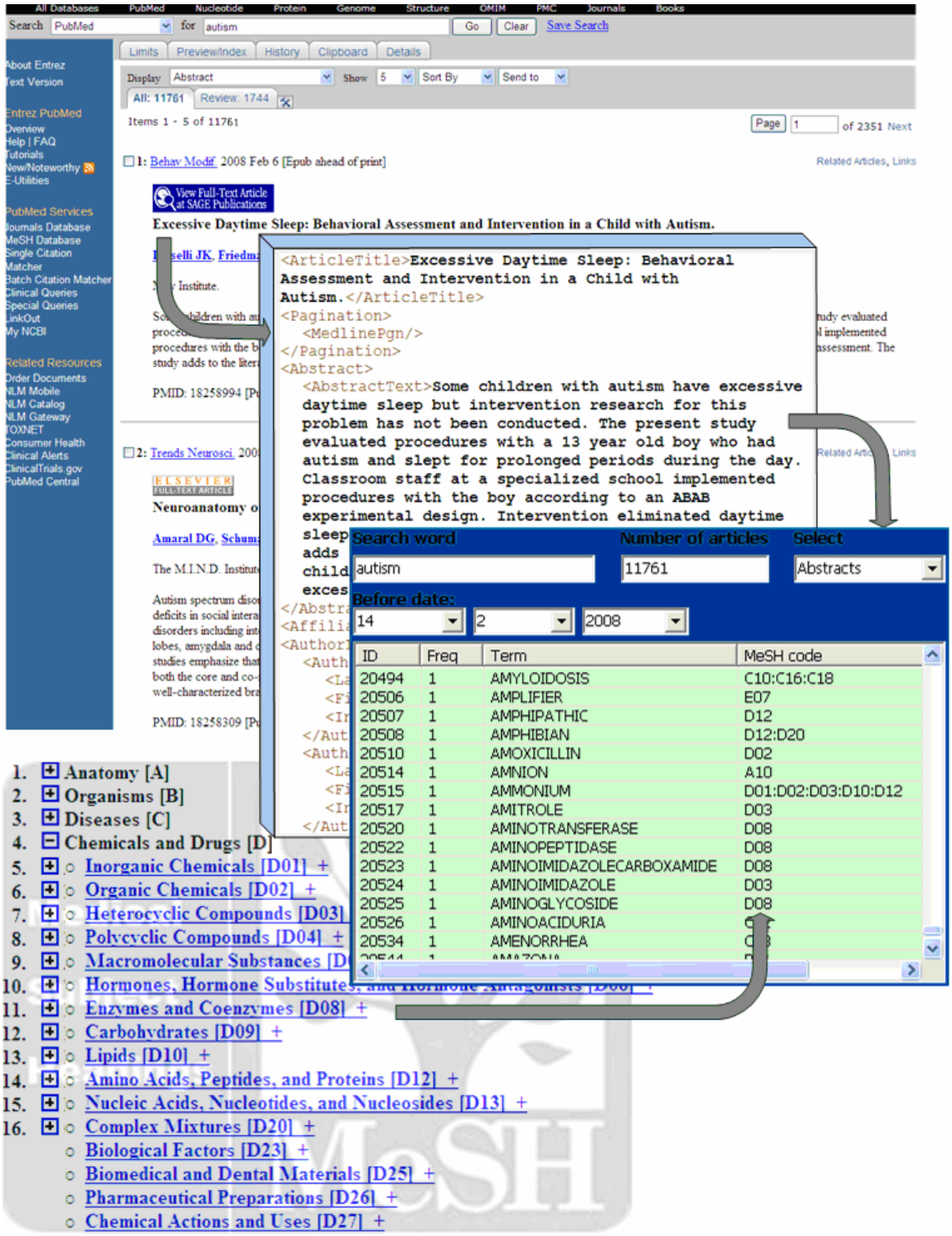


Figure 31: *The Ra step*. Overview of rare terms detection within the RaJoLink's literature-based knowledge discovery approach.

In our experiments we considered only the *D12* second-level category from the 2008 MeSH tree structure, i.e. *Amino Acids, Peptides, and Proteins* (Nelson et al., 2001) because our research has been driven by neurological concepts, which we observed when constructing ontologies on autism documents.

Thus we chose meaningful rare terms belonging to amino acids, peptides, and proteins vocabulary, as in our experimental case the words: *lactoylglutathione*, *synaptophysin*, and *calcium channels*.

In step *Jo*, we collected the MEDLINE abstracts of articles about lactoylglutathione, synaptophysin and calcium channels literature, respectively. We searched for the terms that the 3 text files have in common and thus singled out joint terms for the literatures on the 3 rare terms. From several joint terms that were found automatically, *calcineurin*, a protein phosphatase that is widely present in mammalian brain (Rusnak and Mertz, 2000), was chosen for further investigation.

We began the final step of our method by retrieving abstracts of articles on autism as well as articles on calcineurin from MEDLINE database. In the combined set of literature on autism and literature on calcineurin, we were looking for exceptions within each of the two subgroups of literature, i.e. the calineurin articles among autism major groups of articles, and vice versa.

From the similarity graphs that we drew with OntoGen we could quickly notice which documents are semantically strongly related to each of our research domains, autism and calcineurin, respectively, because they were clearly positioned on the two opposite sides of the similarity curve. However, regarding our goal of looking for relations between our two domains of research, the most prominent examples from the input dataset should be positioned on those graph sides, where the autism articles lay near the calcineurin articles. Therefore we focused our attention on the groups of the calcineurin-autism articles that were positioned in the vicinities according to their similarity. In this way we obtained pairs of calcineurin-autism articles containing terms with similar meanings. We used such terms as a hypothetical conjunct of calcineurin and autism domain. As the candidate hypotheses for calcineurin and autism relationship we found thirteen pairs of MEDLINE articles that, when put together, could connect the two categories, autism and calcineurin, respectively.

According to the expert evaluation, the experimental results resulted in uncovered relations could present a contribution towards better understanding of autism. The reapplication of the RaJoLink method on a restricted set of records mentioning both, autism and fragile X, resulted in some other findings relevant for autism research. More concrete, NF-kappaB was identified as a joint term with potential role in autism.

In fact, it should be mentioned that a full table with identified pairs of related articles proved to be very useful in our dialog with the domain expert since it guided the discussion very efficiently towards new ideas for further investigations. More concretely, the suggestion was to have a closer look at the significance of the fragile X protein loss in autism as reported by Huber and colleagues (Huber et al., 2002). This evaluation significantly helped us in reducing the hypothesis space. It encouraged us to further mine the data on autism in its particular relation to the fragile X. We did it by reapplying the RaJoLink method, this time on a restricted set of articles that dealt with both, autism and fragile X, as described in the following sections.

9.2 Experimental results

9.2.1 Autism and calcineurin relationship

In 1967 Cheung discovered the existence of a protein activator of cyclic nucleotide phosphodiesterase in mammalian brains (Cheung, 1967). Wang and Desai described this modulator binding protein in the year 1977 as a factor that inhibits the cyclic nucleotide phosphodiesterase by the Ca2+-dependent protein modulator (Wang and Desai, 1977).

Klee, Crouch and Krinks gave this inhibitory protein the name *calcineurin* for its Ca2+-dependent mechanism and because of the initial findings, which demonstrated this protein to be specific for the nervous system (Klee et al., 1979). Klee and colleagues determined the calcineurin composition of two polypeptide chains: of a catalytic subunit, calcineurin A and regulatory subunit, calcineurin B. They also indicated that calcineurin is a Ca2+-binding protein with a high affinity for Ca2+ even in the presence of physiological concentrations of Mg2+. Therefore they suggested the calcineurin function in the control of Ca2+-dependent processes in the brain, where Ca2+ has a crucial role in various neuronal processes, such as the biosynthesis and release of neurotransmitters from synaptic vesicles into the synaptic cleft (Klee et al., 1979). Already in 1980, biochemical and immunocytochemical studies associated calcineurin with neuronal elements at postsynaptic sites within neuronal somata and dendrites (Wood et al., 1980).

Distinct studies investigating the mechanisms involved in the inhibition of cell signalling led to groundbreaking discovery that calcineurin is a common target of the immunosuppressant drugs cyclophilin-cyclosporin A and FKBP-FK506 complexes (Liu et al., 1991). Calcineurin inhibition (Figure 32) is

currently practiced to achieve successful immunosuppression in patients after organ transplantation and in treating several other medical problems (Steinbach et al., 2007).

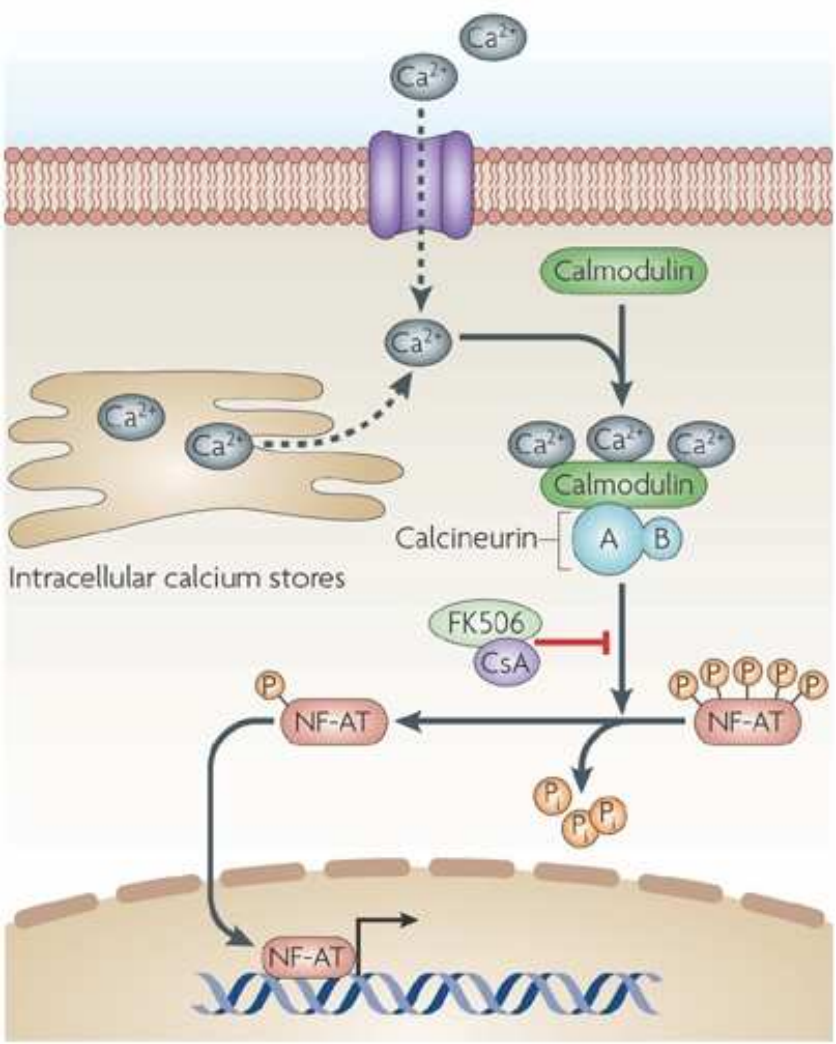


Figure 32: *The calcineurin signalling pathway in T-cells.* Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Microbiology (Steinbach et al., 2007), copyright (2007).

Since the perspective of a domain can be easily captured with ontologies that are based on corpus of text documents, we decided to review the critical points of current calcineurin knowledge through the construction of ontologies. To represent the calcineurin domain in such a data model and to review the concepts inspected in the calcineurin literature, we used OntoGen. We constructed ontologies from abstracts of the calcineurin literature that we retrieved from MEDLINE.

Of 6,290 citations identified on September 9, 2008 by the PubMed search engine, 5,886 were selected for our calcineurin study because their MEDLINE abstracts were available at the time of accession. Figure 33 shows the ontology constructed from the mentioned 5,886 abstracts of calcineurin articles identified through the MEDLINE search. This calcineurin domain ontology consists of five concepts with their most important sub-concepts and their relationships within the domain. It resulted as good domain ontology as it helped us to quickly identify the concept structure of the calcineurin literature and to comprehend the roles that calcineurin plays in various biological processes.

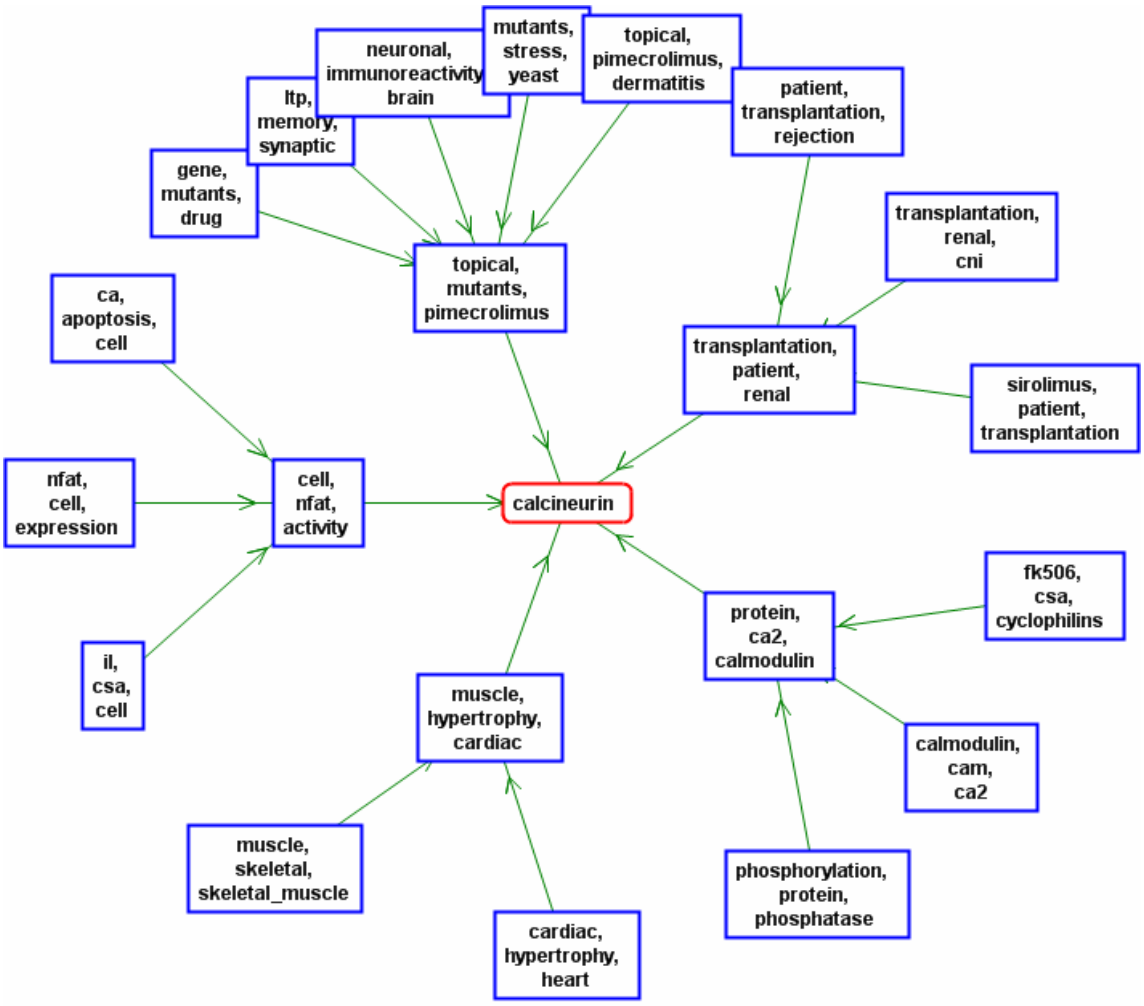


Figure 33: *A two-level calcineurin ontology*. Concepts are named with the keywords suggested by OntoGen.

In our experiments with the combined set of literature on autism and literature on calcineurin, some articles on calcineurin were found in the subgroup of articles on autism according to the semantic similarity measure. Similarly, there were some articles on autism in the subgroup of articles on calcineurin. Such exceptions led us to terms *Bcl-2*, *calmodulin*, *synaptic plasticity*, and ten other linking terms between the literature on autism and the literature on calcineurin. Results are demonstrated according to the Swanson's ABC model in a Venn diagram (Figure 34) and as pairs of MEDLINE articles that link specific autism findings to the calcineurin observations (Table 9).

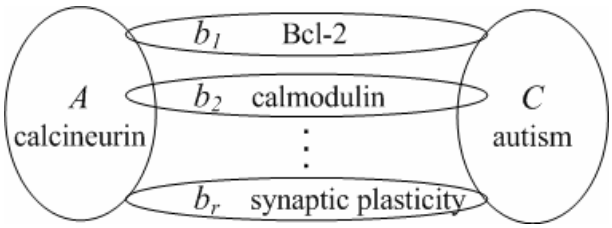


Figure 34: *Venn diagram of arguments (b_i) found as connection between the scientific literature on autism (C) and the scientific literature on calcineurin (A).*

Table 9: *Hypotheses for autism and calcineurin relationship*. Pairs of MEDLINE articles that connect some autism findings on the one hand to the specific calcineurin observations on the other hand.

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.
Roesler et al. (2006) reported about a translocation in the GRPR gene, the mammalian *bombesin*-like peptide gastrin-releasing peptide, being associated with autism.	Corral et al. (2007) showed that *bombesin* promotes the activation of the nuclear factor of activated T cells (NFAT) through a Ca(2+)/calcineurin-linked pathway.
Huber et al. (2002) showed evidences about an important functional role of fragile X protein, an identified cause of autism, in regulating activity-dependent *synaptic plasticity* in the brain.	Winder and Sweatt (2001) described the critical role of protein phosphatase 1, protein phosphatase 2A and calcineurin in the activity-dependent alterations of *synaptic plasticity*.
Belmonte et al. (2004) reviewed neuropathological studies of cerebral cortex in autism indicating abnormal *synaptic* and columnar structure and neuronal migration defects.	Chen et al. (2003) reported about the decrease in protein ubiquitination in synaptosomes and in nonneuronal cells that may play role in the regulation of *synaptic* function by a calcineurin antagonist FK506.
Vorstman et al. (2006) stated that autism spectrum disorders and subthreshold autistic symptoms are common in children with *22q11.2 deletion syndrome*.	Sivagnanasundaram et al. (2007) examined the differential expression of genes mapping to human chromosome 22q11.2 in *22q11.2 deletion syndrome* and found the decreased expression of calmodulin 1 encoding a calcium-dependent protein involved in the calmodulin-calcineurin regulated pathway, which is implicated in learning and memory.
Mouridsen et al. (2007) observed two autoimmune conditions associated with infantile autism: *ulcerative colitis* in mothers and *type 1 diabetes* in fathers of children.	Winter and Schatz (2003) listed immunosuppression by calcineurin inhibitors as one of the promising strategies for intervention in autoimmune *type 1 diabetes mellitus*. Besides this, Shih et al. (2008) suggested the calcium-calcineurin/NFAT pathway as a novel therapeutic target for *ulcerative colitis*.
Román (2007) proposed that morphological brain changes in autism may be produced by *maternal hypothyroxinemia* resulting in low triiodothyronine in the fetal brain during pregnancy.	Sinha et al. (1992) found that calcineurin was compromised in young progeny when they investigated the *maternal hypothyroxinemia* effect during pregnancy on brain of young progeny.
Omura (2006) published the results of measurements of *asbestos* accumulation where relatively high levels of asbestos were found in autism.	Li et al. (2002) investigated the role of reactive oxygen species, by *asbestos*, in activation of nuclear factor of activated T cells (NFAT). They found that pre-treatment of cells with cyclosporin A, a pharmacological inhibitor of calcineurin, blocked asbestos-induced NFAT activation.
Thornton (2006) argued that artificially generated *electromagnetic radiation* may play and important role in the mirror neuron dysfunction associated with autism.	Manikonda et al. (2007) indicated that the exposure to the extremely low frequency *electromagnetic fields* caused increased activities of calcineurin in rat hippocampal regions.
Bernard et al. (2008) revealed that adults with autism showed impaired performance on the tests of *working memory*.	Runyan et al. (2005) illustrated how the inhibition of calcium activated phosphatase calcineurin causes impaired *working memory*.

9.2.2 Autism and NF-kappaB relationship

When verifying hypotheses for the autism-calcineurin relationship by comparing pairs of MEDLINE articles, the medical expert in our team called our attention to the calcineurin role in synaptic plasticity and neuronal activities. Since the impaired synaptic plasticity was reported also in fragile X syndrome that is one of the genetic causes of autism (Irwin et al., 2002) we decided to analyse the autism literature that deals with fragile X in more detail. With the goal to discover unsuspected associations between pieces of knowledge about autism and fragile X, we retrieved articles from MEDLINE that contain information about autism and that at the same time talk about the fragile X. We found 41 articles with their entire text published in the PubMed Central, which served as our input file of data on autism and fragile X. As in the case of our literature mining on pure autism articles, we used them in the open discovery process for the identification of those terms that rarely appeared in the MEDLINE documents collected in our input dataset.

When searching the word frequency statistics we concentrated our attention on listed terms that appeared only in one document from the input dataset. In this way we chose some of these rare terms for the following text mining on the smallest pieces of autism and fragile X knowledge. The following three were chosen based on background knowledge: *BDNF* (brain-derived neurotrophic factor), *bicuculline*, and *c-Fos*. By searching in MEDLINE articles that treat each of the three selected terms domains, we constructed three separate ontologies. Afterwards, we searched the combined files of BDNF, bicuculline and c-Fos articles to find some interesting words that the listed domains have in common as joint terms. We found several promising terms belonging to three of the domains. One of such terms, which we found in the intersection of the three domains, was the term *NF-kappaB*. Figure 35 illustrates how the resulting joint term was obtained for the *autism+fragile_X* domain.

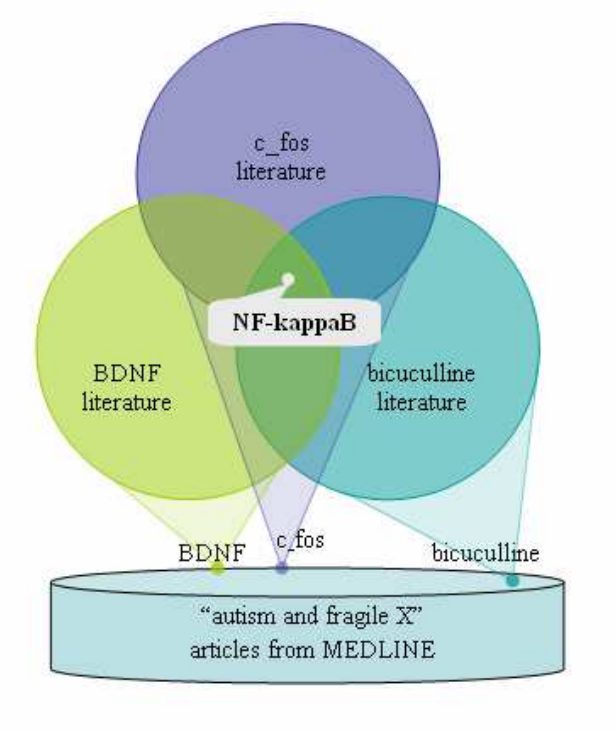


Figure 35: *Experimental results obtained on autism+fragile_X domain*.

The choice of NF-kappaB as a joint term proved to be the right decision, as the appearance of a recent study sponsored by the U.S. National Institute of Mental Health has shown (National Institutes of Health Clinical Center, 2007).

NF-kappaB is a transcription factor that was first discovered by Sen and Baltimore in 1986 through its interaction with the immunoglobulin enhancer sequences (Sen and Baltimore, 1986). They showed that this protein complex has a binding activity specific for the immunoglobulin kappa light chain enhancer sequence. They referred to the binding site for this nuclear factor as the B site and therefore called the factor NF-κB (nuclear factor-kappa B).

To provide tracks of the commonly used terms and to capture the knowledge in the domain of NF-kappaB we generated ontologies like we did in our studies of autism and calcineurin domains. We considered abstracts of MEDLINE articles as a good starting point for knowledge modelling. Therefore we constructed ontologies such as the one in Figure 36 from the 30,893 abstracts of NF-kappaB articles identified through the MEDLINE search on October 17, 2008. The presented NF-kappaB domain ontology consists of four main concepts of the domain and their most important subconcepts.

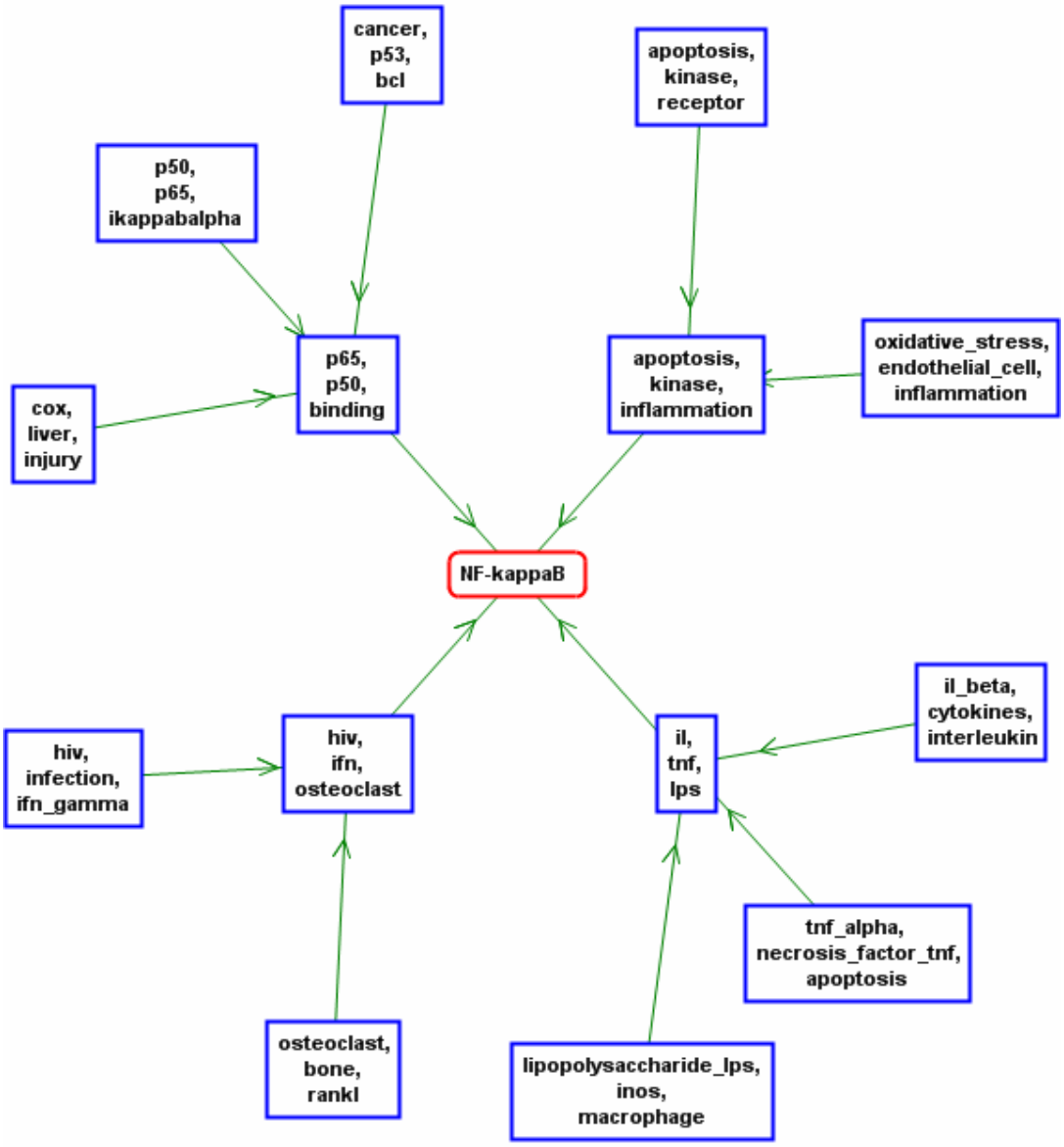


Figure 36: *A two-level ontology that captures the view of NF-kappaB domain.* Concepts are named with the keywords suggested by OntoGen.

To test our hypothesis about the connection between autism and NF-kappaB we analyzed the combined set of abstracts of 9,365 articles on autism and 30,893 articles on NF-kappaB. From the similarity graphs it is easy to identify the documents that are semantically strongly associated with one of the research domains, which were autism and NF-kappaB in our case, because they are clearly positioned on the two opposite sides of the similarity line that represents documents on a graph. Moreover, with the similarity graphs it is simple to detect the documents, which are exceptions within other groups of documents (Figure 37) and that are for this reason very interesting for further investigation.

As we were looking for relations between our two domains of research, namely the autism and the NF-

kappaB domain we focused our attention on those pairs of the NF-kappaB-autism articles that could be localized in the neighbourhood according to their content similarity. Therefore we graphically analyzed the combined input dataset to find articles positioned on those graph sides, where the autism and the NF-kappaB articles lay very near to each other.

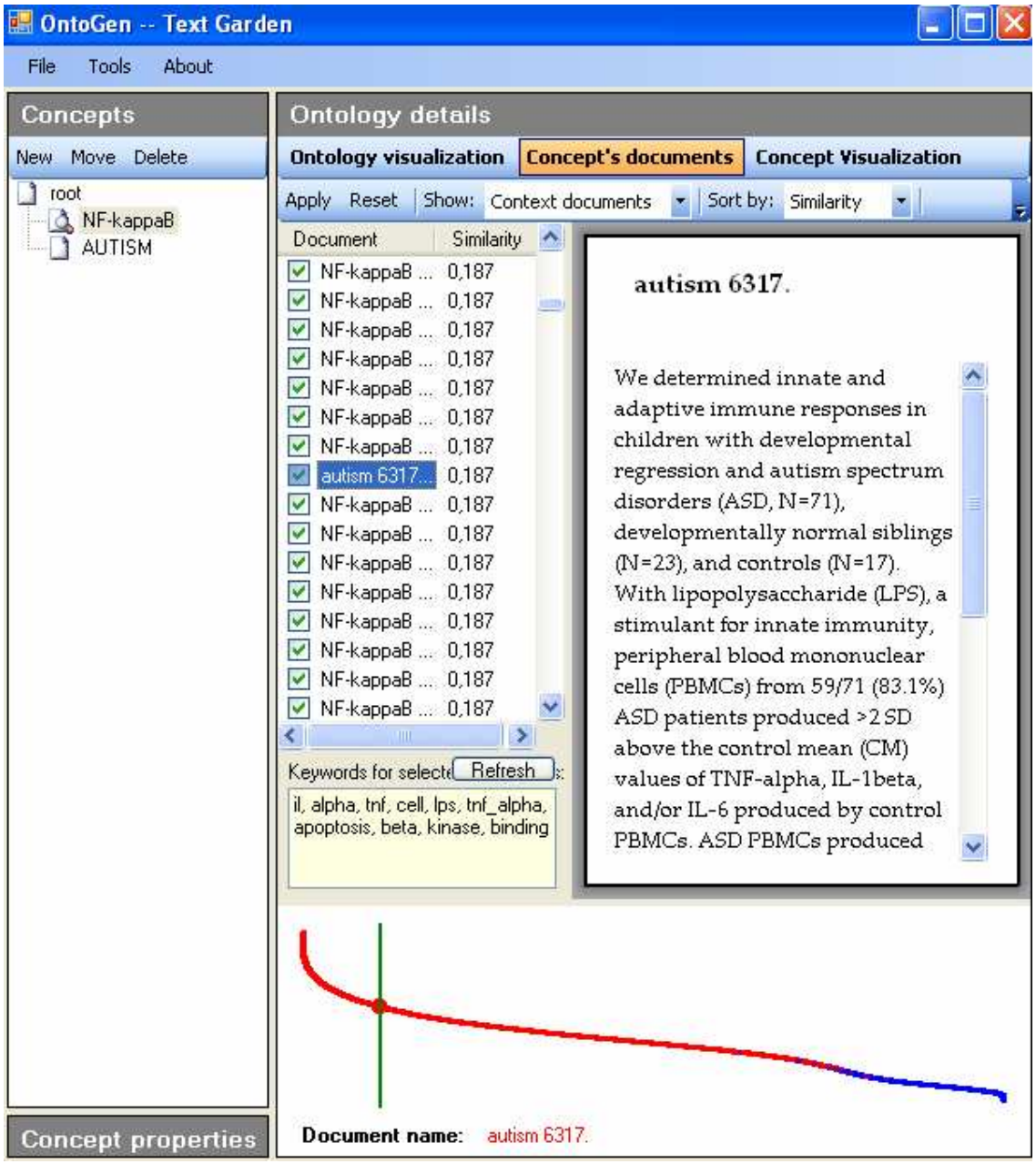


Figure 37: *OntoGen's similarity graph of a combined set of autism and NF-kappaB articles' abstracts*. The central part shows documents that belong to the currently selected NF-kappaB main topic. The article on autism is marked as an exception within the selected topic and positioned among the NF-kappaB context documents.

For a given hypotheses of NF-kappaB and autism relationship we found pairs of MEDLINE articles that could connect the domain of autism with the knowledge gained through the studies of the transcription factor NF-kappaB. In fact, according to the semantic similarity measure we identified some articles on NF-kappaB in the subgroup of articles on autism. In addition, there were also some articles on autism in the subgroup of articles on NF-kappaB. Such exceptions directed us towards the linking terms *Bcl-2*, *cytokines*, *MCP-1*, *oxidative stress* and other meaningful linking terms between the literature on autism and the literature on NF-kappaB. In Table 8 we present ten of the specified pairs of MEDLINE articles that make logical connections between the specific autism observations and the NF-kappaB findings.

Table 10: *Hypotheses for autism and NF-kappaB relationship*. Pairs of MEDLINE articles that connect specific autism findings on the one hand to the NF-kappaB observations on the other hand.

Autism literature	NF-kappaB literature
Araghi-Niknam and Fatemi (2003) showed reduction of *Bcl-2*, an important marker of apoptosis, in frontal, parietal and cerebellar cortices of autistic individuals.	Mattson (2005) reported in his review that activation of NF-kappaB in neurons can promote their survival by inducing the expression of genes encoding antiapoptotic proteins such as *Bcl-2* and the antioxidant enzyme Mn-superoxide dismutase.
Vargas et al. (2005) reported altered *cytokine* expression profiles in brain tissues and cerebrospinal fluid of patients with autism.	Ahn and Aggarwal (2005) reported that on activation NF-kappaB regulates the expression of almost 400 different genes, which include enzymes, *cytokines* (such as TNF, IL-1, IL-6, IL-8, and chemokines), adhesion molecules, cell cycle regulatory molecules, viral proteins, and angiogenic factors.
Vargas et al. (2005) also indicated that macrophage chemoattractant protein *MCP-1* and tumor growth factor-beta1 were the most prevalent cytokines in brain tissues from autistic patients.	Thibeault et al. (2001) showed that *MCP-1* gene is expressed within particular populations of cells in response to inflammatory molecules that employ NF-kappaB as intracellular signaling mechanism.
Ming et al. (2005) reported about the increased urinary excretion of an *oxidative stress* biomarker - 8-iso-PGF2alpha in autism.	Zou and Crews (2006) reported about increase in NF-kappaB DNA binding following *oxidative stress* neurotoxicity.
Yoo et al. (2008) observed statistically significant associations between polymorphisms of PTGS2, the gene encoding *Cyclooxygenase-2* and autism spectrum disorders.	Lee et al. (2004) elucidated the role of spinal NF-kappaB in the *Cyclooxygenase-2* upregulation and pain hypersensitivity following peripheral inflammation.
Ma et al. (2007) performed a genome-wide linkage analysis on 26 extended autism families and found significant linkage to *chromosome 12q14*.	Balaci et al. (2007) mentioned *chromosome 12q14* as a region of IRAK-M gene, which is a nf-kappaB-mediated, negative regulator of the Toll-like receptor/IL-1R pathways.
Steele et al. (2007) demonstrated *spatial memory* deficits in high-functioning individuals with autism, particularly as tasks required heavier demands on working memory .	Denis-Donini et al. (2008) highlighted the function of NF-kappaB in hippocampal neurogenesis and in short-term *spatial memory*.
Jyonouchi et al. (2005) revealed intrinsic defects of *innate immune responses* in children with autism spectrum disorders and gastrointestinal symptoms.	Thomas et al. (2005) confirmed that NF-kappaB has a crucial and multifaceted role in *innate immune responses*.
Grigorenko et al. (2008) identified *macrophage migration inhibitor factor*, which is an upstream regulator of innate immunity as a possible susceptibility gene for autism spectrum disorders.	Gore et al. (2008) showed that *macrophage migration inhibitor factor* regulates subsequent adaptive immune responses by initiating a signalling cascade that activates NF-kappaB.
Johnson and Malow (2008) highlighted frequent sleep problems among children with autism, such as *obstructive sleep apnoea*.	Yamauchi et al. (2006) identified significantly greater activation of NF-kappaB, which occurred in *obstructive sleep apnoea*.

The expert's comment to these findings was as follows: "It is thought that autism could result from an interaction between genetic and environmental factors with an oxidative stress and immunological disorders as potential mechanisms linking the two (Belmonte et al., 2004; Ming et al., 2005). Both of the mechanisms are related to NF-kappaB as the result of our analysis.

The activation of the transcriptional factor NF-kappaB (Figure 38) was shown to prevent neuronal apoptosis in various cell cultures and in vivo models (Mattson, 2005). Oxidative stress and elevation of intracellular calcium levels are particularly important inducers of NF-kappaB activation. In addition, various other genes are responsive to the activation of the NF-kappaB, including those for cytokines. In this way the NF-kappaB can be involved in the complex linkage between the immune system and autism (Belmonte et al., 2004; Vargas et al., 2005). Thus, according to our analysis one possible point of convergence between "oxidative stress" and "immunological disorder" paradigm in autism is NF-kappaB (Macedoni-Lukšič, 2007).

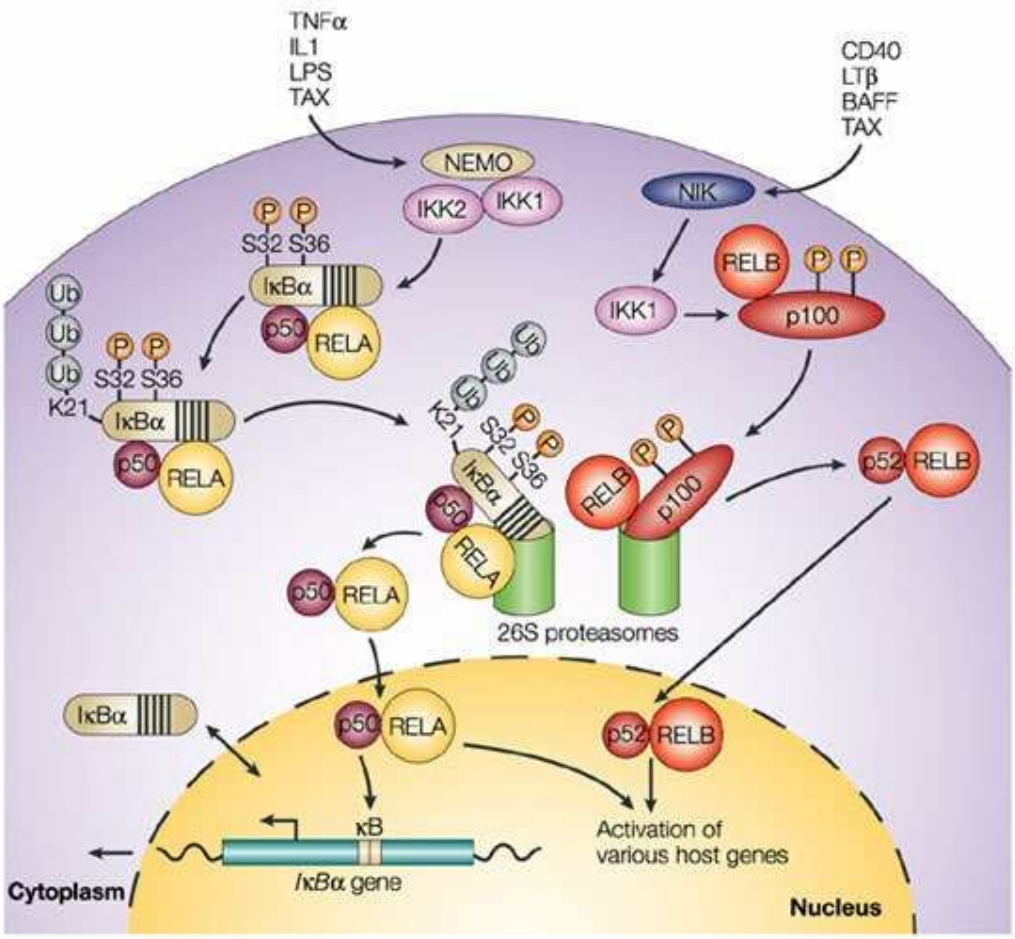


Figure 38: *The NF-kappaB activation pathway.* NF-kappaB is mainly sequestered in the cytoplasm bound to inhibitory IkappaB-alpha proteins. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology (Chen and Greene, 2004), copyright (2004).

10 Evaluation of the RaJoLink Method

The evaluation of the RaJoLink method, presented in this chapter is carried out by considering two experimental cases. First, we evaluate the proposed method in terms of its contribution to understanding of autism. In the second part, we evaluate it in terms of the human effort required for the method execution. With this regard, we investigate the ability of the RaJoLink system to detect the relationship between migraine and magnesium, which is regarded as a gold standard for the literature-based discovery.

10.1 Contribution to understanding of autism

The qualitative evaluation of the generated hypothesis that calcineurin may play an important role in autism was performed by a medical expert. She confirmed the results, which drew attention to interesting connections between two well developed, but not sufficiently connected fields. Her short evaluation was first presented in Urbančič et al. (2007). In particular, she justified this statement by the recent calcineurin studies, indicating that calcineurin participates in intracellular signalling pathways that regulate synaptic plasticity and neuronal activities (Qiu et al., 2006). In addition, she commented that an impaired synaptic plasticity is thought to be also a consequence of the lack of FMR1 protein in fragile X syndrome, which is one of the identified causes of autism (Irwin et al., 2002).

At the time of conducting our experiments, no direct evidence of calcineurin role in the autism phenomena has been reported on the internet yet. Therefore, investigations of the calcineurin role in autism would be of great interest. However, we have found an article, which has been published recently in Molecular Psychiatry (Liu et al., 2007), reporting the evidence of the significant associations of a calcineurin isoform located at the chromosome 8p21.3 region and the schizophrenia subgroup of patients with deficits of the sustained attention and the executive functioning. Although very young autistic children do not demonstrate specific executive dysfunctions, it remains unclear whether executive function deficits present in individuals with autism are the consequence of living with autism or of the difficulties in complex information processing and other non-executive function tasks (Yerys et al., 2007). Besides this, the executive function deficits have been reported as the familial prevalence in first-degree relatives of autistic children (Filipek et al., 2000). Additionally, many very young children with autism have difficulties maintaining their attention to externally imposed tasks (Garretson et al., 1990). Therefore, investigations of the calcineurin role in neurodevelopmental functionalities and neurological difficulties of autism would be of great interest.

Similarly, also the relation between autism and NF-kappaB, which was confirmed as relevant for the autism research during the course of our study (National Institutes of Health Clinical Center, 2007) would be interesting for further examinations. Currently the U.S. National Institutes of Health Clinical Center are recruiting participants to test the effectiveness of minocycline in treating regressive autism through blockade of NF-kappaB nuclear translocation. In fact, the anti-inflammatory antibiotic minocycline is a potent inhibitor of microglial activation, which reduces inflammation by blocking the nuclear translocation of the proinflammatory transcription factor NF-kappaB (National Institutes of Health Clinical Center, 2007).

To further evaluate the RaJoLink method for the value of candidate link discovery in the autism domain we give a discussion of the ten selected linking results (shown in Table 9) for the calcineurin-autism hypothesis and next also for the linking candidates for the NF-kappaB-autism hypothesis (listed in Table 10). For this purpose, we firstly present the short description of the linking results for the calcineurin-autism relation and give explanation of the logic of each association.

Bcl-2 is a regulatory protein for control of programmed brain cell death. Fatemi and colleagues reported a 34 % to 51 % reduction of Bcl-2 levels in autistic cerebellum compared with controls (Fatemi et al., 2001). Their experiments showed that deregulation of Bcl-2 may result in some of the brain structural and behavioural abnormalities in patients with autism. Erin et al., on the other hand, observed that calcineurin occurred as a complex with Bcl-2 in various regions of rat and mouse brain, in particular during times of cellular stress and damage (Erin et al., 2003).

Bombesin is a gastrin-releasing peptide (GRP) homolog. A translocation in the mammalian bombesin-

like peptide gastrin-releasing peptide was associated with autism (Roesler et al., 2006). Besides, Roestler and colleagues pointed that the GRP receptors in brain areas are involved in regulations of synaptic plasticity and autistic behaviours. Moreover, bombesin promotes the activation of the nuclear factor of activated T cells (NFAT) and these effects are obtained through a Ca(2+)/calcineurin-linked pathway (Corral et al., 2007).

Synaptic plasticity provides the structural and functional basis for the maintenance of the complex neural network in the brain. Many researchers examined the diverse functional consequences of regulating activity-dependent synaptic plasticity. Huber and colleagues showed evidences about an important functional role of fragile X protein, an identified cause of autism, in regulating activity-dependent synaptic plasticity in the brain (Huber et al., 2002). In addition, winder and Sweatt described that calcineurin plays the critical role in the activity-dependent alterations of synaptic plasticity (Winder and Sweatt, 2001).

Synapses are specialized intercellular junctions that require neuron specific processes and proteins. Belmonte and colleagues reviewed neuropathological studies of cerebral cortex in autism and indicated abnormal synaptic and columnar structure and neuronal migration defects in patients with autism (Belmonte et al., 2004). Chen et al. concentrated on the regulation of synaptic function by a calcineurin antagonist FK506 (Chen et al., 2003). In particular, they reported about the decrease in protein ubiquitination in synaptosomes and in nonneuronal cells that may play role in the regulation of synaptic function by the calcineurin antagonist FK506.

22q11.2 deletion syndrome results from a micro deletion on the long arm (q) of chromosome 22. The disorder typically involves developmental disability and frequently results in serious psychopathology. Vorstman and colleagues examined psychopathology in patients with the 22q11.2 deletion syndrome and the intelligence level influence on their psychiatric symptoms. Thereafter they found that in children with 22q11.2 deletion syndrome also autism spectrum disorders and subthreshold autistic symptoms are frequent (Vorstman et al., 2006). Sivagnanasundaram et al. applied microarray technology to identify the molecular changes in the hippocampus that may underlie the cognitive deficits and behavioural problems as a result of the 22q11.2 deletion. Thus they found the decreased expression of calmodulin 1 encoding a calcium-dependent protein involved in the calmodulin-calcineurin regulated pathway, which is implicated in learning and memory (Sivagnanasundaram et al., 2007).

Ulcerative colitis and *type 1 diabetes* are both autoimmune conditions. Mouridsen and colleagues observed these two autoimmune diseases associated with infantile autism: on one hand, ulcerative colitis in mothers and on the other hand, the type 1 diabetes in fathers of children with autism (Mouridsen et al., 2007). Winter and Schatz studied the prevention strategies for type 1 diabetes mellitus. As a result, they listed immunosuppression by calcineurin inhibitors as one of the promising strategies for intervention in autoimmune type 1 diabetes mellitus (Winter and Schatz, 2003). Besides, the calcium-calcineurin/NFAT pathway was suggested as a novel therapeutic target for ulcerative colitis because it was highly associated with disease activity (Shih et al., 2008).

Maternal hypothyroxinemia is defined as thyroxine (T4) concentrations that are low for the stage of pregnancy and thus increase the risk of damage to neurodevelopment of the fetus (Morreale de Escobar et al., 2004). Studies show that the most common causes of in utero hypothyroxinemia include maternal flavonoid ingestion during pregnancy as well as environmental antithyroid contaminants (Román, 2007). Román proposed that morphological brain changes in autism may be produced by maternal hypothyroxinemia resulting in low triiodothyronine in the fetal brain during pregnancy (Román, 2007). In the past, when investigating the maternal hypothyroxinemia effect during pregnancy on brain of young progeny, Sinha and colleagues found that calcineurin was compromised in young progeny as a regulator of neurite elongation (Sinha et al., 1992).

Asbestos is a fibrous silicate, which is generally considered the main agent in causing a variety of lung disorders and occupational diseases. Omura published the results of measurements of asbestos accumulation where relatively high levels of asbestos were found also in autistic patients (Omura, 2006). Therefore, the author suggested asbestos to be a possible cause of autism disorders. On the other hand, Li et al. (2002) investigated the role of reactive oxygen species, by asbestos, in activation of nuclear factor of activated T cells (NFAT). They found that pre-treatment of cells with cyclosporin A, a pharmacological inhibitor of calcineurin, blocked asbestos-induced NFAT activation.

Electromagnetic radiation is the artificially generated disturbance in electro-magnetic space and the most likely source of temporal noise in the environment. It is supposed to interfere with the initial calibration of networks of nerve cells in the brain including the mirror neuron system (Thornton, 2006). Thornton argued that artificially generated electromagnetic radiation may play and important role in the mirror neuron dysfunction associated with autism (Thornton, 2006). Accordingly, the author proposed that a temporal disruption from the environment may have an important impact on the mirror neuron dysfunction observed in autism. Recently, Manikonda and colleagues indicated that the exposure of rats to

the extremely low frequency electromagnetic fields caused increased activities of calcineurin in the hippocampal regions (Manikonda et al., 2007).

Working memory is the active, transient maintenance of information in mind that manipulates several cognitive functions. As a consequence, it is involved in many neurological and psychiatric disorders and this way working memory contributes to the cognitive and behavioural dysfunctions associated with such conditions. Bernard and colleagues examined executive dysfunction in adults with autism (Bernard et al., 2008). Their results revealed that adults with autism showed impaired performance on the tests of working memory. Runyan and colleagues, on the other hand, illustrated how the inhibition of calcium activated phosphatase calcineurin causes impaired working memory and indicated that it is critical for working memory tasks (Runyan et al., 2005).

Similarly to a discussion of the ten selected linking results for the calcineurin-autism hypothesis, we also present the analysis of the candidate links that we listed in Table 10 for the NF-kappaB-autism hypothesis. We provide short explanation for each of them and propose how these links may contribute to understanding of autism.

Bcl-2 is an intracellular membrane protein that is a key cell survival regulator. It prevents cells from undergoing apoptosis in response to a variety of cell death signals. Araghi-Niknam and Fatemi showed the reduction of Bcl-2 in superior frontal and cerebellar cortices of autistic individuals by 38 % and 36 %, respectively when compared to tissues of control subjects (Araghi-Niknam and Fatemi, 2003). Mattson, on the other hand reported that activation of NF-kappaB in neurons can promote their survival by inducing the expression of genes encoding antiapoptotic proteins such as Bcl-2 (Mattson, 2005).

Cytokines are small soluble proteins that are produced in response to an antigen and function as mediators and regulators of immunity, inflammation, and haematopoiesis. Vargas and colleagues reported altered cytokine expression profiles in brain tissues and cerebrospinal fluid of patients with autism Vargas et al. (2005). Ahn and Aggarwal specified that on activation NF-kappaB regulates the expression of almost 400 different genes, which include also the inflammatory cytokines, such as TNF, IL-1, IL-6, IL-8, and chemokines (Ahn and Aggarwal, 2005).

MCP-1 is a potent macrophage chemoattractant protein. Vargas and colleagues indicated that MCP-1 and tumor growth factor-beta1 were the most prevalent cytokines in brain tissues from autistic patients (Vargas et al., 2005). Besides, Thibeault and colleagues showed a marked increase in MCP-1 gene expression within particular populations of cells in response to inflammatory molecules that employ NF-kappaB as intracellular signaling mechanism (Thibeault et al., 2001).

Oxidative stress is related to an imbalance between the production of reactive oxygen-derived species and the antioxidant defences. Ming and colleagues reported about the significantly hither urinary excretion of an oxidative stress biomarker - 8-iso-PGF2alpha in autistic patients (Ming et al., 2005). Apart from them, Zou and Crews reported about increase in NF-kappaB DNA binding as a result of the oxidative stress neurotoxicity (Zou and Crews, 2006).

Cyclooxygenase-2 is an inducible enzyme that contributes to the neuroplasticity and the neuropathology of the central nervous system (Yoo et al., 2008). Yoo and colleagues observed statistically significant associations between polymorphisms of PTGS2, the gene encoding Cyclooxygenase-2 and autism spectrum disorders. On the other hand, NF-kappaB regulates the expressions of Cyclooxygenase-2. As an example, Lee and colleagues elucidated the role of spinal NF-kappaB in the Cyclooxygenase-2 upregulation and pain hypersensitivity following peripheral inflammation (Lee et al., 2004).

Chromosome 12q14 indicates one of the regions of the long arm (q) of chromosome 12. Ma and colleagues performed a genome-wide linkage analysis on 26 extended autism families using a high-density single-nucleotide polymorphism genotyping assay and found significant linkage to chromosome 12q14 (Ma et al., 2007). Moreovver, they found this linkage significantly enhanced in the families with only male autistic patients, what suggested a significant gender-specific involvement of the chromosome 12q in the aetiology of autism. Balaci and colleagues mentioned chromosome 12q14 as a region of IRAK-M gene, which is a NF-kappaB-mediated, negative regulator of the Toll-like receptor/IL-1R pathways (Balaci et al., 2007).

Spatial memory typically refers to the capacity of keeping information in mind across trials within a given retention session associated with discriminative spatial localization of reinforcement such as food (de Oliveira and Nakamura-Palacios, 2003). Steele and colleagues demonstrated spatial memory deficits in high-functioning individuals with autism, particularly as tasks required heavier demands on working memory (Steele et al., 2007). On the other side, Denis-Donini and her colleagues observed a defect in short-term spatial memory performance in NF-kappaB p50-deficient mice and this way highlighted the function of NF-kappaB in hippocampal neurogenesis and in short-term spatial memory (Denis-Donini et al., 2008).

Innate immune responses are part of an evolutionarily conserved system of defence that is critically involved in the detection of invading pathogens and the induction of primary adaptive immune responses. Jyonouchi and colleagues revealed intrinsic defects of innate immune responses in children with autism spectrum disorders and gastrointestinal symptoms that were assessed by measuring production of proinflammatory and counter-regulatory cytokines (Jyonouchi et al., 2005). Additionally, it was confirmed that NF-kappaB has a crucial and multifaceted role (e.g., the role of NF-kappaB translocation when it was required for production of the proinflammatory mediators) in innate immune responses (Thomas et al., 2005).

Macrophage migration inhibitor factor is an upstream regulator of innate immunity. Grigorenko and colleagues identified macrophage migration inhibitor factor as a possible susceptibility gene for autism spectrum disorders (Grigorenko et al., 2008). Besides, they found that among family members the patients with autism exhibited higher concentrations of plasma macrophage migration inhibitor factor than their unaffected siblings. Furthermore, Gore and colleagues showed that in B lymphocytes macrophage migration inhibitor factor regulates subsequent adaptive immune responses with initiating a signalling cascade that activates NF-kappaB (Gore et al., 2008).

Obstructive sleep apnoea means interruption in breathing during the sleep hours. Johnson and Malow highlighted frequent sleep problems among children with autism, such as obstructive sleep apnoea (Johnson and Malow, 2008). Separately, Yamauchi and colleagues identified significantly greater activation of NF-kappaB, which occurred in obstructive sleep apnoea (Yamauchi et al. 2006). Taking these two observations together, we can therefore suppose that the activation of NF-kappaB such as that occurs with sleep-disordered breathing may be related also to the pathogenesis of autism spectrum disorders.

Although no direct assessment of calcineurin and NF-kappaB involvement in the phenomenon of autism has been published yet, we have been able to identify significant links between calcineurin and autism literature and similarly between NF-kappaB and autism literature by combining the articles from two domains in a single set of literature and searching for semantic similarities among documents from such combined input set. However, we observe that it is difficult to isolate important domain concepts, such as neurological aspects, without proper background knowledge. Therefore, by interacting with a subject expert, the entire process of knowledge discovery can benefit in terms of speed and guidance towards meaningful solutions. Especially in the cases where no lexical classification is available for the researched domain, the domain expert can provide the necessary background knowledge for the identification of meaningful results.

In order to get a straightforward medical confirmation of our hypotheses generated in the autism domain, additional expert investigations are needed. Since this may take some time, we wanted to check the capabilities of RaJoLink also in another way, more suitable for an immediate evaluation. To answer today if RaJoLink has the potential to reveal future discoveries without waiting these discoveries to actually be confirmed, we decided to make an experiment as it was done in the past. Therefore, we wanted to investigate whether by having all evidence that was available at a certain time in the past, could RaJoLink point to discoveries that were not known at that particular moment, but were confirmed some years later? To answer this question we carried out an experiment in the migraine domain, simulating the Swanson's migraine – magnesium discovery as follows in the next section.

10.2 Required human effort

Due to an enormous quantity of articles available on-line, literature-based knowledge discovery has become a very time consuming and laborious task. The RaJoLink method is intended to support experts on their search for new discoveries. It covers both, open discovery and closed discovery processes. In both processes it tends to reduce the required human effort.

Swanson stated that in open discovery processes success depends entirely on the knowledge and ingenuity of the searcher (Swanson, 1990). The aim of RaJoLink is to reduce the search space, thus making the task easier for the searcher. At the same time, by focusing on rare terms the system identifies the candidates that are most likely to lead towards meaningful unpublished relations. This way, the system automatically produces intermediate results. Search space is further reduced by human choices of rare terms and joint terms. In addition, human involvement in these steps assures that search process concentrates on those parts of the search space that are interesting and meaningful for a subject expert. Based on these strategies, RaJoLink is designed to make the expert's involvement friendly and more efficient.

To quantitatively evaluate the performance of the RaJoLink system we decided to apply the RaJoLink

method to yet another important application domain. Therefore, we replicated the early Swanson's migraine-magnesium experiment that represents a gold standard for the literature-based discovery. Moreover, besides the magnesium, we estimated also the rest of the potential discoveries that the RaJoLink system generated in the open discovery process in the migraine domain. This way we performed a more complete evaluation of the RaJoLink method, which gives additional information about the method's capability of detecting interesting terms in large text collections.

Like Swanson in his original study of the migraine literature (Swanson, 1988) we used titles as input for our literature-based discovery. However, we excluded from the analysis the article that mentioned both migraine and magnesium in their title (Vosgerau, 1973) although it was published before the Swanson's discovery of the migraine-magnesium connection. However, this article was not analysed by Swanson, most probably because it was written in German. This way we proved the originality of the magnesium discovery in our migraine experiment.

With the automatic support of the RaJoLink system, we performed the experiment on the entire set of the MEDLINE titles of articles that were published before 1988 and that we retrieved with the search of the phrase: *migraine NOT magnesium*. As a result we got 6,135 titles of the MEDLINE articles that we analysed according to the RaJoLink method to identify interesting discoveries in connection with migraine.

With the goal to support the domain expert during the choices of the potentially relevant rare terms in the open discovery process, we performed the analysis of the migraine-magnesium experimental results by observing which MeSH categories are significant for new discoveries. By estimating the number of rare terms within a particular MeSH category that led to the discovery of magnesium as a joint term, we obtained the density distribution of MeSH categories. Based on the number of rare terms that resulted as relevant for the hypotheses generation in the step *Jo*, we can observe the influence of a particular MeSH category to the achievement of results. The findings of this analysis are demonstrated in Figure 39.

In the open discovery process of the migraine experiment we were able to obtain another three important discoveries related to migraine, besides magnesium. In fact, among the interesting joint terms, we identified also the terms:

- *interferon* that denotes proteins, which are produced by the cells of the immune system,
- *interleukin* that is a type of signalling molecules called cytokines,
- *tnf* that stands for tumour necrosis factor.

Figure 39 presents the results for the four joint terms, namely the terms *interferon*, *interleukin*, *magnesium*, and *tnf* that we examined in more detail during the open discovery process of the migraine experiment. For the sake of this experiment we automatically retrieved from the MEDLINE database 1000 titles of articles for each of the selected rare terms identified in the literature about migraine. Again, the PubMed search and text analysis was performed on the articles published before 1988 with the support of the RaJoLink system. We calculated the document frequency statistics for all potentially relevant rare terms. The MeSH category rankings were computed separately for each of the selected joint terms: *interferon*, *interleukin*, *magnesium*, and *tnf*.

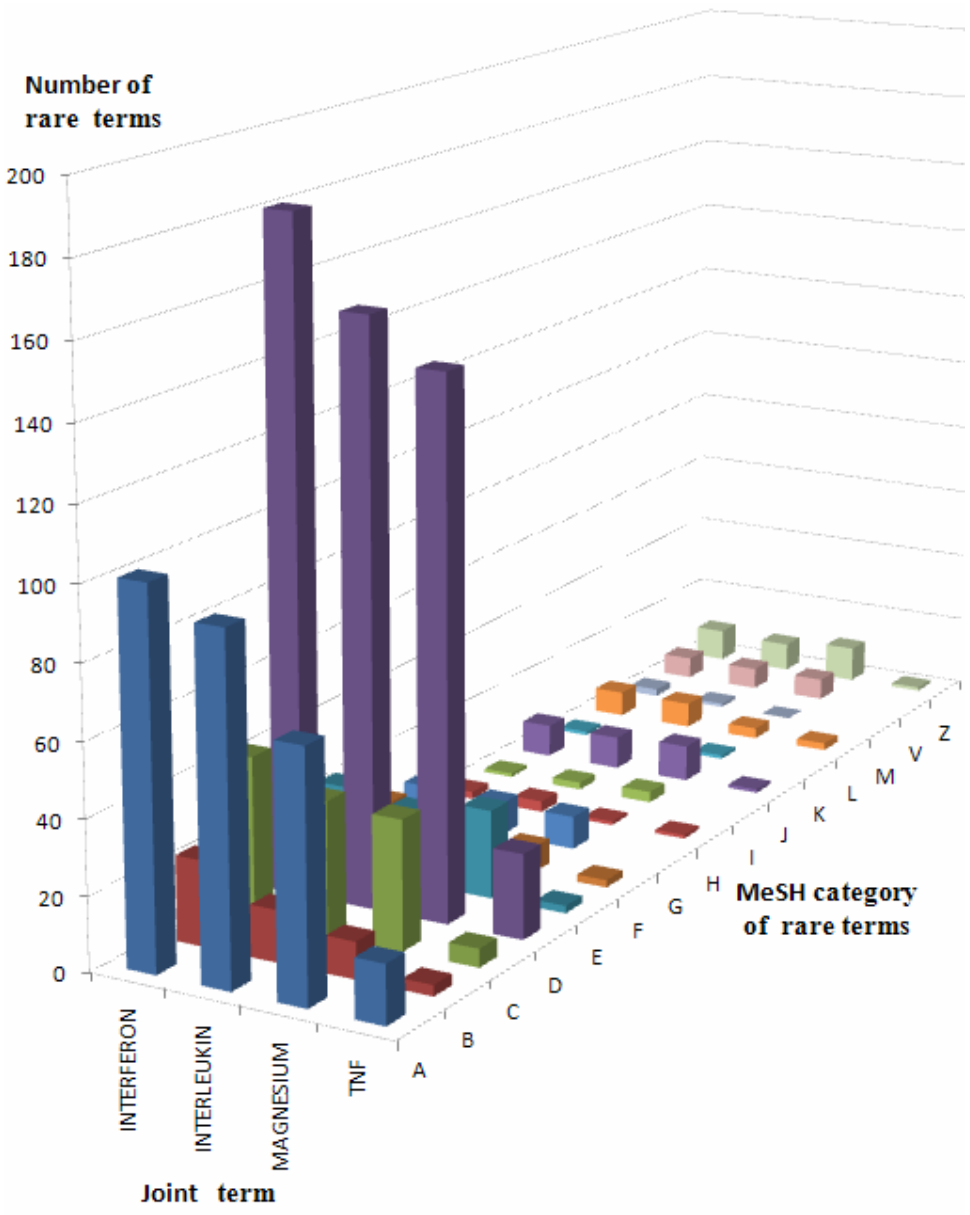


Figure 39: *The results of the RaJoLink method applied to the domain of migraine*. Besides magnesium, another 3 discoveries were obtained with the RaJoLink system, namely the joint terms: interferon, interleukin and tnf.

As shown in Figure 39, the rare terms from the MeSH category D – Chemicals and Drugs were the most prominent in the migraine literature because the largest number of terms that led to the discovery of the four joint terms were from the category of Chemicals and Drugs. The second category of terms that frequently led to the four discoveries was the MeSH category A – Anatomy. The third significant MeSH category for each of the analyzed joint terms was the category C – Diseases.

Since the detection of rare terms and their grouping together according to their MeSH categories was done automatically without the help of a medical expert, these ranks represent the reliability and the objectivity of the calculated statistics. Therefore they provide an objective view to the evaluation of the MeSH categories and their role in choosing the most perspective rare terms for new discoveries in a biomedical domain, such as migraine.

The discoveries of the interferon's, interleukin's and tnf's role and their underlying mechanisms

involved in the migraine headaches have been recently reported in 16 scientific articles about interferon, in 39 studies of interleukin and in 29 scientific studies of tumour necrosis factor (Source: MEDLINE search, December 2008).

On the other hand, with the text analysis of the articles that were published in MEDLINE before 1988, we succeeded in identifying interferon, interleukin and tnf as potential discoveries in relation to migraine although no article about their involvement in migraine can be found in MEDLINE before the year 1988. In fact, tnf and interleukin were first associated with migraine in articles accessible through MEDLINE in 1990 and 1991, respectively (Covelli et al., 1990; Covelli et al., 1991). The connection between interferon and migraine was established even later, starting in 1995 according to articles published in MEDLINE (Detrey-Morel et al., 1995).

In the migraine domain, we performed another analysis of rare terms by focusing on their MeSH categories. With the RaJoLink analysis of the 6,135 titles of the MEDLINE articles about migraine that were published before 1988, we identified 4,669 different terms. To obtain a particular estimate of the rank of rare terms with their document frequency equal to 1, which led to the discovery of magnesium, we analyzed the rare terms together with their MeSH categories in which the magnesium occurred as a target term in the Jo step. We analysed their rank separately for each of the MeSH categories.

We analysed the document frequency for each unique rare term that the RaJoLink system identified in the collection of titles of articles on migraine. For the purpose of the RaJoLink's evaluation we considered in detail the meaningful rare terms with regard to their MeSH categories and excluded those terms that can be treated as stop words (e.g. *symbol*, *type*, and *subgroup*). Figure 40 shows the rare terms together with their document frequencies calculated for the titles that included the term magnesium.

We used the document frequencies of terms to rank the rare terms according to their likely relevance for linking the starting concept with the target term *magnesium*. As we already stated, the document frequency of a term is the number of documents in which that particular term occurs. As a result of this analysis, we obtained a top ranked list of terms with their MeSH categories (Figure 40). These top ranked categories can be considered by human experts as the most interesting ones for generation of hypothesis in the biomedical domains of research. Accordingly, the terms and their MeSH categories with a higher ranking will more likely draw the attention of a human expert and thus they provide an expert-guided discovery model.

In the MeSH category C – Diseases, magnesium receives the top rank for the term *hypocalcemia* that appeared in 56 titles of the observed articles together with the term *magnesium*. However, the most of the rare terms that appeared also in more than five titles of documents about magnesium belong to the MeSH category D – Chemicals and Drugs. Such rare terms that we identified in the titles of the articles about migraine are *phosphorus*, *oxalate*, *salicylate*, *potassium*, *parathormone*, and *terbutaline*. This means that the migraine–magnesium hypothesis is very plausible if we take more terms from the MeSH category D – Chemicals and Drugs. Besides, this category was found as the most prominent in our migraine experiment also for the discoveries of interferon, interleukin and tnf.

As demonstrated in Figure 40, we managed to identify several rare terms in our training datasets that have led us to discover magnesium in relation with migraine. These results indicate that with the RaJoLink method it is possible to discover knowledge from different domains, which are usually treated separately (e.g., in genetic, epidemiologic, clinical or other scientific literature). Rare terms are able to indicate if there exists a joint term (i.e. magnesium), which can be linked to the domain under study (i.e. migraine in this case). Thus the identification of valuable rare terms can enrich the knowledge discovery process.

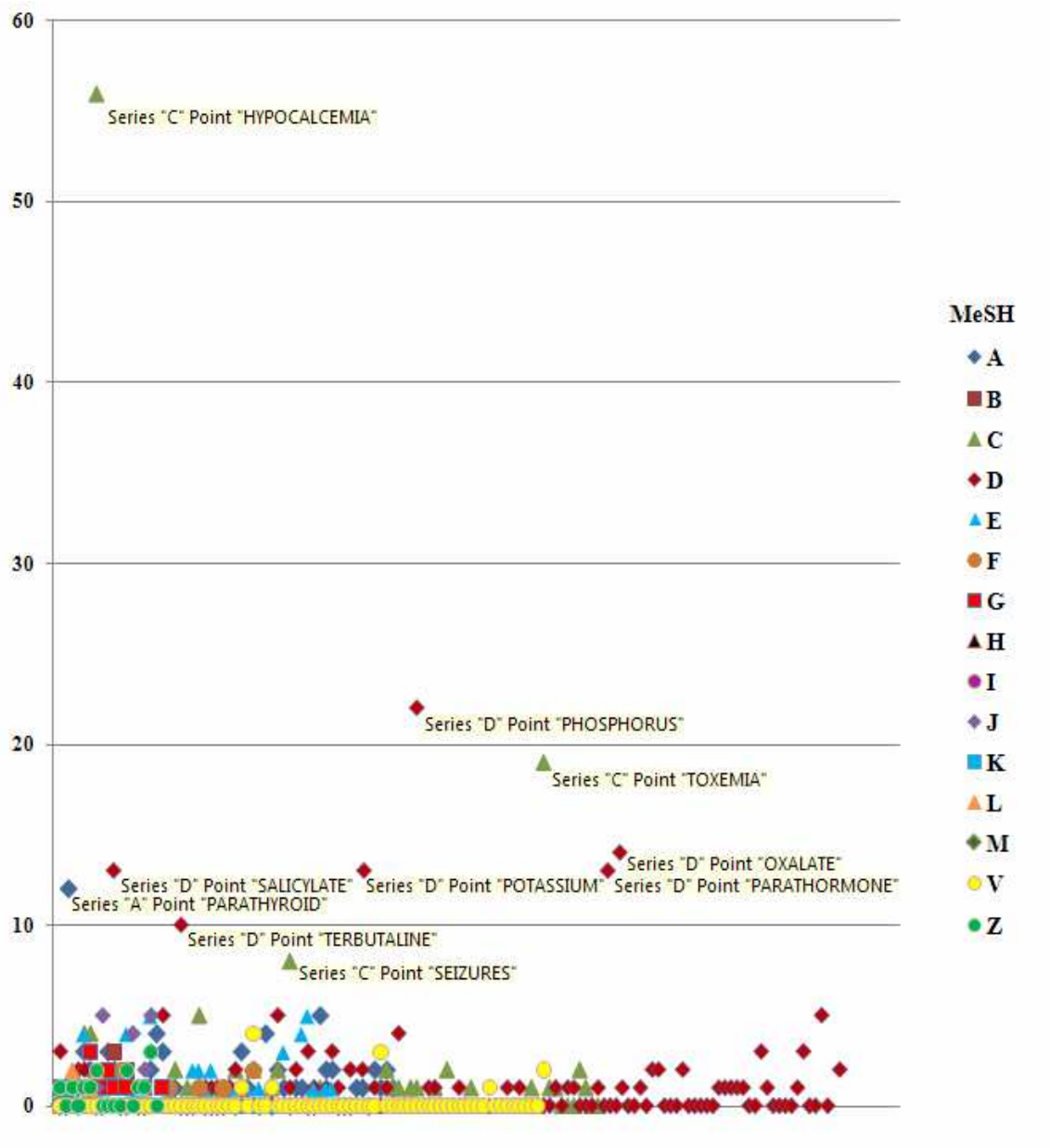


Figure 40: *The document frequencies of rare terms that appeared together with the term magnesium in the titles of MEDLINE documents published before 1988.* The rare terms (e.g. the term *seizures*) that appeared in more than five titles of articles together with the term *magnesium* are marked with their series name (i.e. the MeSH category) and the point name (i.e. the observed rare term).

The text mining process normally identifies a huge number of rare terms in a document collection. This leads to a question of how a subset of terms could be determined that would form a valid statistical criterion for selection of a relevant subset according to the knowledge domain. With our experimental results we confirmed that MeSH categorisation can effectively help the human expert with identifying the relevant target terms. Besides this, we improved the RaJoLink's performance by providing further quantitative evaluation of results as described in the following section.

10.3 Improvements of the RaJoLink's performance based on the evaluation results

The evaluation of the RaJoLink method in terms of the human effort that would be required for the detection of the relationship between migraine and magnesium showed us that, although MeSH filtering is very helpful in identifying interesting terms from the huge list of results, the method needs further improvements in the restriction of the potential candidates' lists for each discovery task.

Following the migraine-magnesium evaluation results we improved RaJoLink in step *Ra* and in step *Jo*. Firstly, in step *Ra*, we provided the possibility to automatically rank rare terms according to the number of MEDLINE articles, which contain a rare term together with the starting term *c*. This way any higher co-occurrence of a rare term with the starting term *c* indicates the already explored connection, which is not interesting from a novelty point of view. Therefore, the user should regard only those terms, which have the lowest value for the *C & Rare term* threshold (i.e. starting from the value 1). The list of rare terms can be further filtered by MeSH categories as can be seen in Figure 41, where from all 4,669 different terms identified in the literature on migraine (domain *C*), only the truly rarest terms from the MeSH categories A – Anatomy, C – Diseases and D – Chemicals and Drugs are displayed. Such a list helps experts evaluate the possible terms from the starting literature in order to determine the ones which represent the very rare but meaningful terms in the domain literature under research.

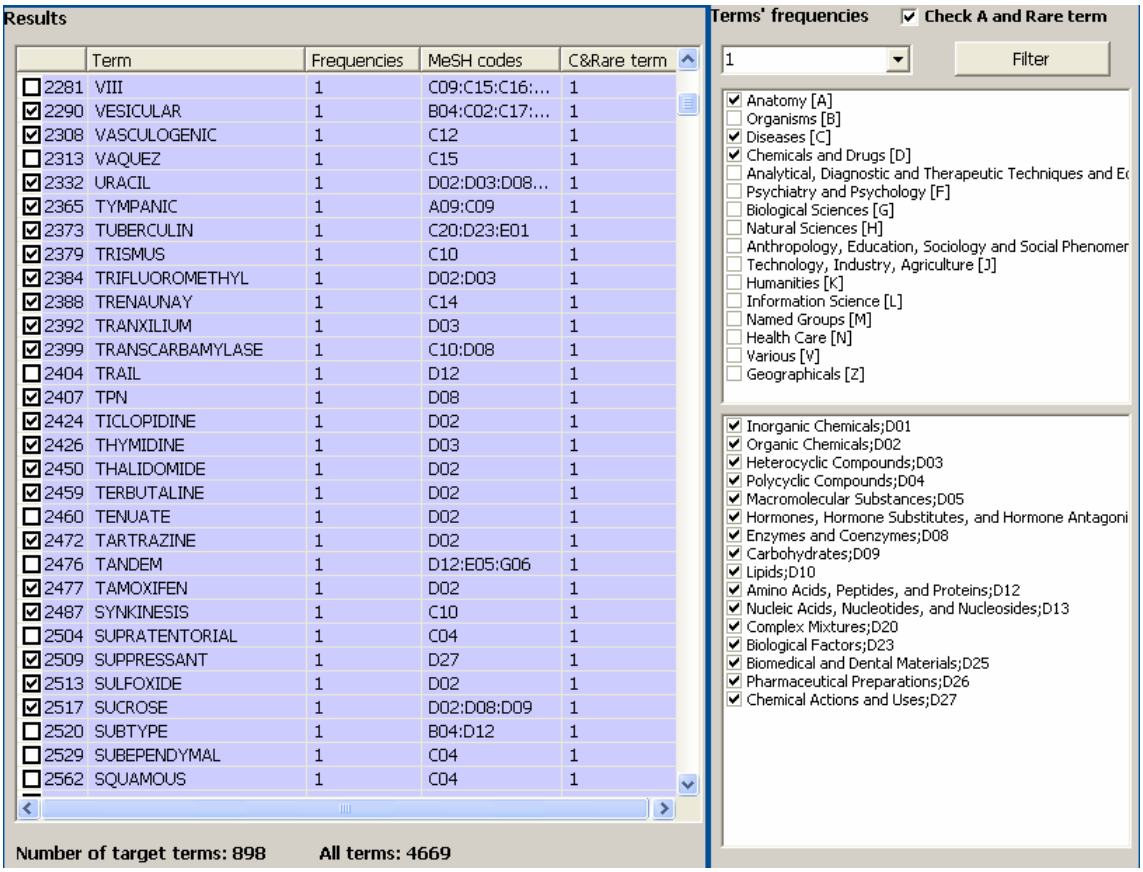


Figure 41: *A screenshot of RaJoLink showing filtered results in step Ra for the migraine-magnesium experiment*. The selected rare terms from the literature about migraine are chosen for further analysis.

When we replicated the analysis of the 6,135 titles of the MEDLINE articles about migraine that were published before 1988, we were able to restrict the target rare terms from the all identified 4,669 different terms to only 898 terms from the MeSH categories A, C and D and furthermore to only 100 meaningful rarest terms that each appeared just in one MEDLINE article on migraine. Nine of these rarest terms from the migraine literature led to the discovery of magnesium in step *Jo*, with the analysis of 100 titles for each of the 100 selected rare terms. These discoveries have proven that RaJoLink is capable of automatically

identifying meaningful connections in the open discovery process without requiring the user to manually prune the non-relevant candidates from the lists of text mining results.

Similarly, in step *Jo*, we included the automatic ranking of the candidate joint terms, where RaJoLink checks each candidate joint term *a* whether it appears together with the starting term *c* in the MEDLINE articles. If a candidate joint term *a* is found in any article together with the starting term *c*, the value for the *A & C* assessment becomes greater than 0, which means that a selected joint term *a* cannot be regarded as a novel discovery in the domain under research. Accordingly, the A & C assessment solves the second problem of pruning those candidates for joint terms *a* that have been already connected with the starting term *c* in the MEDLINE articles. When we applied the novel RaJoLink functionality on the migraine-magnesium experiment, it resulted in fewer but more reliable results for candidate joint terms.

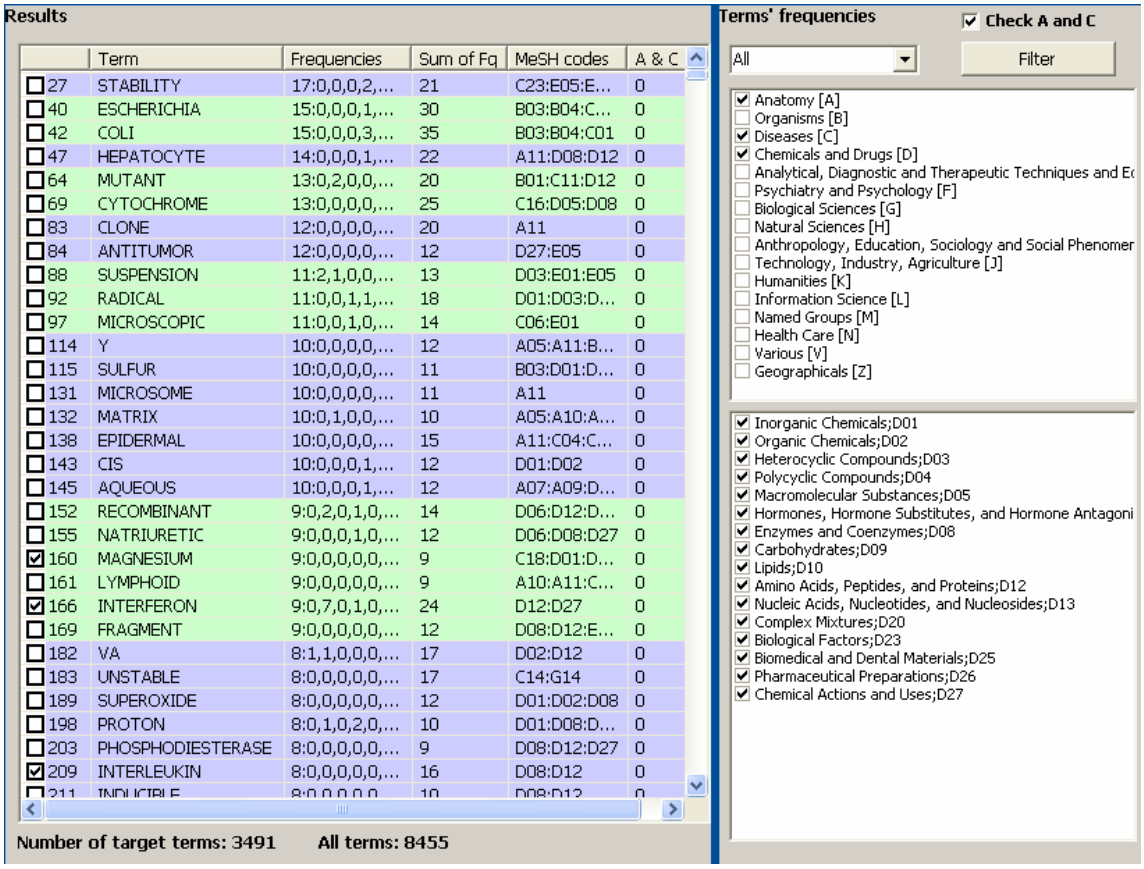


Figure 42: *A screenshot of RaJoLink showing filtered results for candidate joint terms*. The terms magnesium, interferon and interleukin appear on the top 30 list of candidate joint terms.

When replicating the migraine-magnesium experiment, the terms magnesium, interferon and interleukin appeared on the top 30 list of candidate joint terms among the 8,455 terms identified in step *Jo* as shown in the screenshot of RaJoLink (Figure 42). Our replication thus shows that increasing the quality of RaJoLink features improves the rank of relevant novel discoveries and is therefore effective in automatically identifying novel relationships in the published biomedical literature.

11 Conclusions

The knowledge gathered by various specialised sciences throughout the digital era has resulted in large volumes of data and complex data interrelationships. Extracting useful knowledge in the form of uncovered relations can therefore be very time consuming. To support biomedical experts in their knowledge discovery process, we have developed a literature mining method called RaJoLink that uncovers hidden relations from large sets of scientific articles in a given domain. The method implements the Swanson's ABC model approach for generating hypotheses in a new innovative way without knowing the target concept in advance, which is discovered later within the process. The RaJoLink method searches for logically connected pieces of literature on rare terms identified in the literature on a given phenomenon under investigation (e.g., disease). This way it supports human expert in the process of generating and testing hypotheses in the domain under study. The method is named RaJoLink after its key procedural elements, which are: rare terms, joint terms and linking terms. Consequently, the entire RaJoLink's approach consists of three principal steps, *Ra*, *Jo* and *Link*.

In step *Ra*, a specified number (set by user as a parameter value) of interesting rare terms in literature about the phenomenon *C* under investigation are identified. In step *Jo*, all available articles about the selected rare terms are inspected and interesting joint terms that appear in the intersection of the literatures about rare terms are identified. One of them is selected as the candidate for *A*. In step *Link*, linking terms *b*, which bridge literature about *A* and literature about *C*, are searched for. Relations between *A* and *C* are established via *AB* and *BC* relations. Evaluation of pairs (*AB*, *BC*) as support for potential hypotheses about the relation between *A* and *C* is carried out by the domain expert.

In the described RaJoLink method, the knowledge discovery process starts with the open discovery, in which hypotheses are not known in advance. One of the main advantages of RaJoLink therefore, lays in the support of the open discovery processes by the innovative use of rare terms from the problem domain literature to guide the generation of new hypotheses. Accordingly, the crucial step of the method consists of selecting rare terms that are identified in the literature *C*. The intuition behind this research idea was that the rarer a term is in the domain literature, the higher is the probability to encounter observations that represent something unexpected that may lead to creative discovery of new knowledge. This way we managed to employ rarity as a principle and means to find new interesting pieces of knowledge that were previously available in the dispersed literature and could be linked together.

The unique contribution of the RaJoLink method and a fundamental difference from the previously proposed models of the open discovery approach lies in the rarity principle that we apply to the open literature-based discovery. In fact, we use rare terms identified in the literature *C* to guide the search for new hypotheses. To this end, we have applied the rarity principle together with the notion of bisociation. In fact, the context-crossing connections, called bisociations, are often needed for creative, innovative discoveries (Koestler, 1964).

Bisociative relationships can only be discovered on the basis of a sufficiently large and diverse underlying corpus of information. In our case this corpus are MEDLINE papers. The larger the corpus is, the more likely it is to contain bisociative relationships. The RaJoLink approach has the potential for bisociative relation discovery as it allows switching between contexts (papers from different areas) by exploring rare terms in the intersection between contexts (Petrič et al., 2007; Urbančič et al., 2007; Petrič et al., 2009).

Besides this, we contributed an innovative approach also to the closed discovery process. The closed discovery process in RaJoLink is based on the outliers' detection in the content similarity graphs. In fact, having two disjoint literatures *A* and *C*, we automatically search for linking terms that are mentioned in both, the literature *A* as well as in the literature *C*. Pairs of documents with the same linking terms are subject to closer inspection in order to find out whether by putting statements about a linking term in these two articles together supports the hypothesis about a meaningful relation between previously disjoint literatures. In this manner, our search for linking terms is done in a semi-automated way that reduces manual work and efficiently points to meaningful relations between the concepts *a* and *c*.

The huge sets of biomedical articles that augment the space of possible hypotheses for problem solutions represent a significant challenge in the field of biomedical discoveries based on literature. We

have carried out a series of investigations about the differences between titles, abstracts and bodies of texts (Petrič et al., 2006a; Cestnik et al., 2007). For this purpose we constructed and analysed several ontologies and drew comparisons between them. Articles about autism from the MEDLINE database served as our testbed. First, we provided an overview of autism phenotype and aetiology research. Then we presented ontology construction approaches and our studies on documents about autism. At this stage, we compared and evaluated the obtained ontologies. Ontologies built with OntoGen helped us to substantially speed up the process of reviewing and understanding the complex and heterogeneous spectrum of scientific articles about autism. The main concepts of autism phenomena as they result from the first level of our ontology models are: genetics, autism treatments, epidemiology, neurobiology and environmental factors. Significant verification of the resulted ontology construction can be obtained from the recent state of autism research represented by Zerhouni (2004) that summarizes the main scientific activities of autism research in the major areas of epidemiology, genetics, neurobiology, environmental influence and specific methods for treating autism.

We used ontologies in the knowledge exchanging context as representations of autism theories. Our observations show that ontologies help sharing understanding in a given subject area. Therefore, we propose to include ontologies in knowledge discovery frameworks for facilitating communication among interdisciplinary groups of experts. In particular, ontologies help exchanging views between knowledge engineers and domain experts and thus verifying the understanding of the domain knowledge.

In the comparison of autism ontologies, built with OntoGen, we focused on vocabulary level of results of automatic concepts construction. A graphical presentation of comparison results was 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. Compared to the analysis of matching bodies of texts and abstracts, we observed significantly lower similarity between texts and titles of the related articles, as well as between their abstracts and titles. Articles about genetics are the only fairly important group of documents that apparently use more similar vocabulary in their titles and abstracts and in the entire bodies of their texts. The likely cause for this observation lies in the genetic terminology and in the genetic context itself, which is reasonably specific when compared to other fields of autism research.

We also conducted experiments in which the RaJoLink method was applied to a concrete problem domain (Petrič et al., 2006b; Petrič et al., 2007; Urbančič et al., 2007; Petrič et al., 2009). With our experimental results we wanted to show how connections between rare concepts that appear in distinct contexts of domain under study can lead to innovative discoveries in the open discovery process. When we applied the method to the autism domain, we discovered calcineurin as a joint term in the intersection of the literature about the selected rare concepts that were identified in the autism literature. The results were medically confirmed as relevant for the autism research during the course of our study, and as interesting for further examinations. Similarly, the transcription factor NF-kappaB was recognized as a joint concept.

In the closed discovery process we presented important connections among information embedded in autism and calcineurin literature, as well as relationships between autism and NF-kappaB literature. We discovered such connections by analysing outliers in the published evidence of some autism findings on one hand that coincide with specific calcineurin and NF-kappaB observations on the other hand.

The main motivation for our study of this biomedical domain was the desire to discover connections, crucial for understanding of autism developmental disorders for earlier and more reliable medical diagnosis and intervention strategies for children with autism. As autism incidence is increasing in Slovenia, as well as in the world, and it profoundly affects the patient and his family, every new piece of knowledge that helps better understand this disorder is welcome. However, further research about timing, environmental conditions, maturational differences in brain development, and other determinants of calcineurin and NF-kappaB involvement in autism spectrum disorders is needed for stronger evidence.

Besides this, also the methodological relevance of results is important, since it is difficult to capture all the available information with existing machine learning methods from heterogeneous data available from different sources of various types. Actually, the results of the initial experimental case studies in autism domain suggest that the RaJoLink method can enhance the state-of-the-art methods for the literature-based discovery. To evaluate the RaJoLink method, we investigated its ability to detect also the relationship between migraine and magnesium, which has been frequently used as a gold standard in the literature-based discovery community. Specifically, we investigated two issues: having information about papers that were available in 1988, (1) does RaJoLink find magnesium, and (2) does RaJoLink maybe find something else that has later proved to be connected with migraine. In the migraine documents published before the discovery of the migraine-magnesium relationship, we managed to identify several rare terms that have led us to discover magnesium in relation with migraine. Moreover, besides the magnesium, we were able to obtain another three important discoveries related to migraine, namely the terms *interferon*, *interleukin*, and

tnf. In fact, at the time of the Swanson's experiment, interferon, interleukin and tumour necrosis factor (tnf) were not connected with migraine in MEDLINE articles at all. The connections appeared in MEDLINE some years later.

For improvement of RaJoLink we have suggested further directions for its development. Among them, the automated identification of semantic exceptions in texts, and the prediction of relevant rare terms for analysis in closed discovery process and more guidance to the users are planned in the future. All in all, we hope that the presented work will improve the scientific progress based on knowledge discovery from the scientific literature and will provide support for experts on their way towards new discoveries in biomedical domains.

12 Acknowledgements

The work on this dissertation has been assisted by the encouragement and help of several persons. I am especially grateful to my supervisors, Prof. Dr. Tanja Urbančič and Doc. Dr. Bojan Cestnik for their assistance, suggestions, and advice. I wish to express my sincere appreciation to both of them.

I sincerely thank Dr. Marta Macedoni-Lukšič from University Children's Hospital, University Medical Center of Ljubljana for her guidance and evaluations from the medical point of view, which significantly influenced this work. I also want to acknowledge the University of Nova Gorica for providing the financial support for my doctoral studies.

I am thankful to the Department of Knowledge Technologies at Jožef Stefan Institute. My special thanks go to Prof. Dr. Nada Lavrač for her invaluable suggestions, Prof. Dr. Marko Bohanec for reviews of my work, Blaž Fortuna for his discussions about OntoGen's performance and Dr. Martin Žnidaršič for his practical advice.

I also wish to thank Prof. Dr. Olga Štěpánková and Prof. Dr. Filip Železný from Czech Technical University for their helpful comments and suggestions for improvements of RaJoLink.

Finally, I am particularly grateful to my parents and to my husband for their continuous support and encouragement throughout my studies.

13 References

[1] Aggarwal, C. C.; Yu, P. S. An effective and efficient algorithm for high-dimensional outlier detection. *International Journal on Very Large Data Bases* **14(2)**, pp 211–221 (2005).

[2] Ahn, K. S.; Aggarwal, B. B. Transcription Factor NF-{kappa}B: A Sensor for Smoke and Stress Signals. *Annals of the New York Academy of Sciences* **1056**, pp 218-233 (2005).

[3] Almind, T. C.; Ingwersen, P. Informetric Analyses on the World Wide Web: Methodological Approaches to "WEBOMETRICS". *Journal of Documentation* **53(4)**, pp 404-426 (1997).

[4] Altura BM. Calcium antagonist properties of magnesium: implications for antimigraine actions. *Magnesium* **4(4)**, pp 169-175 (1985).

[5] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition, Text Revision* (Washington, DC, 2000).

[6] Anselin, L.; Sridharan, S.; Gholston, S. Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns. *Social Indicators Research* **82(2)**, pp 287–309 (2007).

[7] Araghi-Niknam, M.; Fatemi, S. H. Levels of Bcl-2 and P53 are altered in superior frontal and cerebellar cortices of autistic subjects. *Cellular and Molecular Neurobiology* **23(6)**, pp 945-952 (2003).

[8] Balaci, L.; Spada, M. C.; Olla, N.; Sole, G.; Loddo, L.; Anedda, F.; Naitza, S.; Zuncheddu, M. A.; Maschio, A.; Altea, D.; Uda, M.; Pilia, S.; Sanna, S.; Masala, M.; Crisponi, L.; Fattori, M.; Devoto, M.; Doratiotto, S.; Rassu, S.; Mereu, S.; Giua, E.; Cadeddu, N. G.; Atzeni, R.; Pelosi, U.; Corrias, A.; Perra, R.; Torrazza, P. L.; Pirina, P.; Ginesu, F.; Marcias, S.; Schintu, M. G.; Del Giacco, G. S.; Manconi, P. E.; Malerba, G.; Bisognin, A.; Trabetti, E.; Boner, A.; Pescollderungg, L.; Pignatti, P. F.; Schlessinger, D.; Cao, A.; Pilia, G. IRAK-M is involved in the pathogenesis of early-onset persistent asthma. *The American Journal of Human Genetics* **80(6)**, pp 1103-1114 (2007).

[9] Barnard, L.; Muldoon, K.; Hasan, R.; O'Brien, G.; Stewart, M. Profiling executive dysfunction in adults with autism and comorbid learning disability. *Autism* **12(2)**, pp 125-141 (2008).

[10] Barnett, V.; Lewis, T. *Outliers in statistical data* (Wiley, New York, 1994).

[11] Bauman, M. L.; Kemper, T. L. Neuroanatomic observations of the brain in autism: a review and future directions. *International Journal of Developmental Neuroscience* **23(2-3)**, pp 183-187 (2005).

[12] Batagelj, V.; Mrvar, A. Density based approaches to network analysis: analysis of Reuters terror news network. In: *Workshop on Link Analysis for Detecting Complex Behavior LinkKDD2003*, (2003).

[13] Batagelj, V.; Mrvar, A. Pajek: a program for large network analysis. *Connections* **21(2)**, pp 47-57 (1998).

[14] Becchetti, L.; Castillo, C.; Donato, D.; Leonardi, S.; Baeza-Yates, R. Link Analysis for web spam detection. *ACM Transactions on the Web* **2(1)**, pp 1-42 (2008).

[15] Becker, K. G.; Hosack, D. A.; Dennis, G. Jr.; Lempicki, R. A.; Bright, T. J.; Cheadle, C.; Engel, J. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* **4**, (2003).

[16] Belmonte, M. K.; Allen, G.; Beckel-Mitchener, A.; Boulanger, L. M.; Carper, R. A.; Webb, S. J. Autism and abnormal development of brain connectivity. *The Journal of Neuroscience* **24(42)**, pp 9228-9231 (2004).

[17] Bethea, T. C.; Sikich, L. Early pharmacological treatment of autism: a rationale for developmental treatment. *Biological Psychiatry* **61(4)**, pp 521-537 (2007).

[18] Bettelheim, B. The empty fortress: Infantile autism and the birth of the self. (Free Press, New York, 1967).

[19] Boughton, D. Paradoxes in science: A new view of rarity. *Science findings of Pacific Northwest Research Station* **35**, (2001).

[20] Brank, J.; Grobelnik, M.; Milić-Frayling, N.; Mladenić, D. Feature selection using support vector

machines. In: Zanisi, A.; Brebbia, C. A.; Ebecken, N. F. F. E.; Melli P. (eds) *Data Mining III*. pp 261-273 (WIT Press, Southampton, Boston, 2002).

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

[22] Carney, R. S. Basing conservation policies for the deep-sea floor on current-diversity concepts: A consideration of rarity. *Biodiversity and Conservation* **6(11)**, pp 1463-1485 (1997).

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

[24] Chen, D.; Müller, H. M.; Sternberg, P. W. Automatic document classification of biological literature. *BMC Bioinformatics* **7**, (2006).

[25] Chen, H.; Polo, S.; Di Fiore, P. P.; De Camilli, P. V. Rapid Ca2+-dependent decrease of protein ubiquitination at synapses. *Proceedings of the National Academy of Sciences of the United States of America* **100(25)**, pp 14908-14913 (2003).

[26] Chen, L. F.; Greene, W. C. Shaping the nuclear action of NF-kappaB. *Nature Reviews Molecular Cell Biology* **5(5)**, pp 392-401 (2004).

[27] Chen, X. Cognitive relevance and chance discovery. In: Abe, A.; Ohsawa, Y. (eds) *Readings in Chance Discovery. Advanced Knowledge International*, pp 169-190 (2005).

[28] Cheung, W.Y. Cyclic 3',5'-nucleotide phosphodiesterase: pronounced stimulation by snake venom. *Biochemical and Biophysical Research Communications* **29(4)**, pp 478-482 (1967).

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

[30] Cohen, A. M.; Yang, J.; Hersh, W. R. A Comparison of Techniques for Classification and Ad Hoc Retrieval of Biomedical Documents. In: *Proceedings of the Fourteenth Annual Text REtrieval Conference, TREC 2005* (Gaithersburg, MD, National Institute for Standards & Technology, 2005).

[31] Corral, R. S.; Iniguez, M. A., Duque, J.; Lopez-Perez, R., Fresno, M. Bombesin induces cyclooxygenase-2 expression through the activation of the nuclear factor of activated T cells and enhances cell migration in Caco-2 colon carcinoma cells. *Oncogene* **26(7)**, pp 958-969 (2007).

[32] Cory, K.A. Discovering hidden analogies in an online humanities database. *Library Trends* **48(1)**, pp 60–71 (1999).

[33] Covelli, V.; Munno, I.; Pellegrino, N. M.; Altamura, M.; Decandia, P.; Marcuccio, C.; Di Venere, A.; Jirillo, E. Are TNF-alpha and IL-1 beta relevant in the pathogenesis of migraine without aura? Review. *Acta Neurologica (Napoli)* **13(2)**, pp 205-211 (1991).

[34] Covelli, V.; Munno, I.; Pellegrino, N. M.; Di Venere, A.; Jirillo, E.; Buscaino, G. A. Exaggerated spontaneous release of tumor necrosis factor-alpha/cachectin in patients with migraine without aura. *Acta Neurologica (Napoli)* **12(4)**, pp 257-263 (1990).

[35] de Oliveira, R. W.; Nakamura-Palacios E. M. Haloperidol increases the disruptive effect of alcohol on spatial working memory in rats: a dopaminergic modulation in the medial prefrontal cortex. *Psychopharmacology* **170(1)**, pp 51-61 (2003).

[36] Denis-Donini, S.; Dellarole, A.; Crociara, P.; Francese, M. T.; Bortolotto, V.; Quadrato, G.; Canonico, P. L.; Orsetti, M.; Ghi, P.; Memo, M.; Bonini, S. A.; Ferrari-Toninelli, G.; Grilli, M. Impaired adult neurogenesis associated with short-term memory defects in NF-kappaB p50-deficient mice. *The Journal of Neuroscience* **28(15)**, pp 3911-3919 (2008).

[37] Detry-Morel, M.; Boschi, A.; Gehenot, M.; Geubel, A. Bilateral transient visual obscurations with headaches during alpha-II interferon therapy: a case report. *European Journal of Ophthalmology* **5(4)**, pp 271-274 (1995).

[38] DeVito, T. J.; Drost, D. J.; Neufeld, R. W.; Rajakumar, N.; Pavlosky, W.; Williamson, P.; Nicolson, R. Evidence for cortical dysfunction in autism: a proton magnetic resonance spectroscopic imaging study. *Biological Psychiatry* **61(4)**, pp 465-473 (2007).

[39] Ellison, A. M.; Agrawal, A. A. The Statistics of Rarity. *Ecology* **86(5)**, pp 1079-1080 (2005).

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

[41] European Communities. The Information Society Technologies (IST) Work Programme, IST Projects Fact Sheets. http://cordis.europa.eu/ (accessed November 2008).

[42] European Network of Ombudspersons for Children (ENOC). Report of the 10th ENOC Annual Meeting. Athens, Greece, 26-28 September 2006. http://www.ombudsnet.org/enoc/meetings/index.asp (accessed November 2006).

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

[44] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* **17(3)**, pp 37-54 (1996).

[45] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge discovery and data mining: towards a unifying framework. In: *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon, 1996).

[46] Feldman, R.; Dagan, I. Knowledge Discovery in Textual Databases (KDT). In: *Proceedings of the 1st International Conference on Knowledge Discovery KDD-95*, pp 112-117 (Montreal, Canada, 1995).

[47] Feldman, R.; Sanger, J. *The Text Mining Handbook: Advanced approaches in analyzing unstructured data* (Cambridge University Press, 2006).

[48] Filipek, P. A.; Accardo, P. J.; Ashwal, S.; Baranek, G. T.; Cook, E. H. Jr.; Dawson, G.; Gordon, B.; Gravel, J. S.; Johnson, C. P.; Kallen, R. J.; Levy, S. E.; Minshew, N. J.; Ozonoff, S.; Prizant, B. M.; Rapin, I.; Rogers, S. J.; Stone, W. L.; Teplin, S. W.; Tuchman, R. F.; Volkmar, F. R. Practice parameter: screening and diagnosis of autism: report of the Quality Standards Subcommittee of the American Academy of Neurology and the Child Neurology Society. *Neurology* **55(4)**, pp 468-479 (2000).

[49] Fortuna, B.; Grobelnik, M.; Mladenić, D. Semi-automatic data-driven ontology construction system. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O.; Musek, J.; Osredkar, M. J.; Kononenko, I.; Novak Škarja, B. (eds) I*S-2006. Proceedings of the 9th International multi-conference Information Society*. pp 223-226 (Ljubljana, Slovenia, 2006).

[50] Frei, C.; Schär, C. Detection probability of trends in rare events: Theory and application to heavy precipitation in the Alpine region. *Journal of Climate* **14(7)**, pp 1568-1584 (2001).

[51] Friedman, C.; Shagina, L.; Lussier, Y.; Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association* **11(5)**, pp 392-402 (2004).

[52] Garretson, H. B.; Fein, D.; Waterhouse, L. Sustained attention in children with autism. *Journal of Autism and Developmental Disorders* **20(1)**, pp 101-114 (1990).

[53] Gennari, J.; Musen, M. A.; Fergerson, R.W.; Grosso, W. E.; Crubezy, M.; Eriksson, H.; Noy, N. F.; Tu, S.W. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development (2002). http://smi.stanford.edu/smi-web/reports/SMI-2002-0943.pdf (accessed December 2006).

[54] Georgiades, S.; Szatmari, P.; Zwaigenbaum, L.; Duku, E.; Bryson, S.; Roberts, W.; Goldberg, J.; Mahoney, W. Structure of the autism symptom phenotype: A proposed multidimensional model. *Journal of Amer Academy of Child & Adolescent Psychiatry* **46(2)**, pp 188-196 (2007).

[55] Giles, C. B.; Wren, J. D. Large-scale directional relationship extraction and resolution. *BMC Bioinformatics* **9 (Suppl 9)**, pp S11 (2008).

[56] Gore, Y.; Starlets, D.; Maharshak, N.; Becker-Herman, S.; Kaneyuki, U.; Leng, L.; Bucala, R.; Shachar, I. Macrophage migration inhibitory factor induces B cell survival by activation of a CD74-CD44 receptor complex. *Journal of Biological Chemistry* **283(5)**, pp 2784-2792 (2008).

[57] Grigorenko, E. L.; Han, S. S.; Yrigollen, C. M.; Leng, L.; Mizue, Y.; Anderson, G. M.; Mulder, E. J.; de Bildt, A.; Minderaa, R. B.; Volkmar, F. R.; Chang, J. T.; Bucala, R. Macrophage migration inhibitory factor and autism spectrum disorders. *Pediatrics* **122(2)**, pp e438-e445 (2008).

[58] Grobelnik, M.; Mladenić, D. Automated knowledge discovery in advanced knowledge management. *Journal of Knowledge Management* **9(5)**, pp 132-149 (2005).

[59] Grobelnik, M.; Mladenić, D. Extracting human expertise from existing ontologies. In: *EU-IST Project IST-2003-506826 SEKT* (2004).

[60] Hearst, M. A. Untangling text data mining. In: Dale, R., (ed.) *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* 3-10 (Morgan Kaufmann Publishers, San Francisco, CA, 1999).

[61] Hirschman, L.; Morgan, A. A.; Yeh, A. S. Rutabaga by any other name: extracting biological names.

Journal of Biomedical Informatics **35(4)**, pp 247-259 (2002).

[62] Hirtz, D.; Thurman, D. J.; Gwinn-Hardy, K.; Mohamed, M.; Chaudhuri, A. R.; Zalutsky, R. How common are the "common" neurologic disorders? *Neurology* **68(5)**, pp 326-337 (2007).

[63] Hollingsworth, B.; Lewin, I.; Tidhar, D. Retrieving Hierarchical Text Structure from Typeset Scientific Articles - a Prerequisite for E-Science Text Mining. In: Cox, S. J. (ed.) *Proceedings of the 4th UK E-Science All Hands Meeting*. 267-273 (Nottingham, 2005).

[64] Hristovski, D.; Friedman, C.; Rindflesch, T. C.; Peterlin, B. Exploiting Semantic Relations for Literature-Based Discovery. *AMIA Annual Symposium Proceedings*, pp 349–353 (2006).

[65] Hristovski, D.; Peterlin, B.; Mitchell, J. A.; Humphrey, S. M. Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics* **74(2-4)**, pp 289-298 (2005).

[66] Huber, K. M.; Gallagher, S. M.; Warren, S. T.; Bear, M. F. Altered synaptic plasticity in a mouse model of fragile X mental retardation. *Proceedings of the National Academy of Sciences of the United States of America* **99(11)**, pp 7746-7750 (2002).

[67] IPCC. *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (2007).

[68] Irwin, S.; Galvez, R.; Weiler, I. J.; Beckel-Mitchener, A.; Greenough, W. Brain structure and the functions of FMR1 protein. In: Hagerman, R. J., Hagerman, P. J. (eds) *Fragile X syndrome*. 191-205 (The Johns Hopkins University Press, Baltimore, 2002).

[69] Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: a review. *ACM Computing Surveys* **31(3)**, pp 264–323 (1999).

[70] Johnson, K. P.; Malow, B.A. Sleep in children with autism spectrum disorders. *Current Treatment Options in Neurology* **10(5)**, pp 350-359 (2008).

[71] Joshi, A.; Undercoffer, J. L. On Data Mining, Semantics, and Intrusion Detection. What to Dig for and Where to Find It. In: Kargupta, H.; Joshi, A.; Sivakumar, K.; Yesha, Y. (eds) *Data mining. Next Generation Challenges and Future Directions*. 437-460 (Menlo Park, California, 2004).

[72] Juršič, M.; Mozetič, I.; Lavrač, N. Learning ripple down rules for efficient lemmatization. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O. (eds) *IS 2007. Proceedings of the 10th International Multiconference Information Society*. pp 206-209 (Ljubljana, Slovenia, 2007).

[73] Jyonouchi, H.; Geng, L.; Ruby, A.; Zimmerman-Bier, B. Dysregulated innate immune responses in young children with autism spectrum disorders: their relationship to gastrointestinal symptoms and dietary intervention. *Neuropsychobiology* **51(2)**, pp 77-85 (2005).

[74] Kanner, L. autistic disturbances of affective contact. *Nervous Child* 2, pp 217-250 (1943).

[75] Klee, C. B.; Crouch, T. H.; Krinks, M. H. Calcineurin: a calcium- and calmodulin-binding protein of the nervous system. *Proceedings of the National Academy of Sciences of the United States of America* **76(12)**, pp 6270-6273 (1979).

[76] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM* **46(5)**, pp 604-632 (1999).

[77] Klin, A.; Volkmar, F. R. *Asperger's Syndrome, Guidelines for Assessment and Diagnosis* (Pittsburgh, Learning Disabilities Association of America, 1995).

[78] Koestler, A. *The act of creation* (MacMillan Company, New York, 1964).

[79] Lavrač, N.; Gamberger, D. Saturation filtering for noise and outlier detection. In: *12th European Conference on Machine Learning / 5th European Conference on Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD-2001. Active learning, database sampling, experimental design: views on instance selection: working notes*. pp 1-4 (Freiburg, Germany, 2001).

[80] Lawson, A. E. What does Galileo's discovery of Jupiter's moons tell us about the process of scientific discovery? *Science Education* **11**, pp 1–24 (2002).

[81] Lazarevic, A.; Kumar, V.; Srivastava, J. Intrusion detection: A survey. In: Kumar, V.; Srivastava, J.; Lazarevic, A. (eds) *Massive Computing, Managing Cyber Threats*. pp 19-80 (Springer, US, 2005).

[82] Lee, K. M.; Kang, B. S.; Lee, H. L.; Son, S. J.; Hwang, S. H.; Kim, D. S.; Park, J. S.; Cho, H. J. Spinal NF-kB activation induces COX-2 upregulation and contributes to inflammatory pain hypersensitivity. *European Journal of Neuroscience* **19(12)**, pp 3375-3381 (2004).

[83] Leung, C. K.-S.; Thulasiram, R. K.; Bondarenko, D. A. An Efficient System for Detecting Outliers from Financial Time Series. In: Bell, D.; Hong, J. (eds) *Flexible and Efficient Information Handling*. pp 190-198 (Springer, Berlin, 2006).

[84] Li, J.; Huang, B.; Shi, X.; Castranova, V.; Vallyathan, V.; Huang, C. Involvement of hydrogen peroxide in asbestos-induced NFAT activation. *Molecular and Cellular Biochemistry* **234-235(1-2)**, pp 161-168 (2002).

[85] Liu, F.; Jenssen, T. K.; Nygaard, V.; Sack, J.; Hovig, E. FigSearch: a figure legend indexing and classification system. *Bioinformatics* **20(16)**, pp 2880-2882 (2004).

[86] Liu, H.; Lieberman, H.; Selker, T. A Model of Textual Affect Sensing using Real-World Knowledge. In: *Proceedings of the 2003 International Conference on Intelligent User Interfaces, IUI 2003*. pp 125-132 (Miami, Florida, USA, 2003).

[87] Lindsay, R. K.; Gordon, M. D. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science and Technology* **50(7)**, pp 574-587 (1999).

[88] Liu, J.; Farmer, J. D. Jr; Lane, W. S.; Friedman, J.; Weissman, I.; Schreiber, S. L. Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. *Cell* **66(4)**, pp 807-815 (1991).

[89] Liu, Y. L.; Fann, C. S.; Liu, C. M.; Chang, C. C.; Yang, W. C.; Hung, S. I.; Yu, S. L.; Hwang, T. J.; Hsieh, M. H.; Liu, C. C.; Tsuang, M. M.; Wu, J. Y.; Jou, Y. S.; Faraone, S. V.; Tsuang, M. T.; Chen, W. J.; Hwu, H. G. More evidence supports the association of PPP3CC with schizophrenia. *Molecular Psychiatry* **12(10)**, pp 1-9 (2007).

[90] Ma, D. Q.; Cuccaro, M. L.; Jaworski, J. M.; Haynes, C. S.; Stephan, D. A.; Parod, J.; Abramson, R. K.; Wright, H. H.; Gilbert, J. R.; Haines, J. L.; Pericak-Vance, M. A. Dissecting the locus heterogeneity of autism: significant linkage to chromosome 12q14. *Molecular Psychiatry* **12(4)**, pp 376-384 (2007).

[91] Macedoni-Lukšič, M. Personal communication (2007).

[92] Maedche, A.; Staab, S. Measuring Similarity between Ontologies. In: *Lecture Notes In Computer Science: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*. pp 251 - 263 (2002).

[93] Magnani, L. Chance discovery and the disembodiment of mind. In: Khosla, R.; Howlett, R. J.; Jain L. C. (eds) *Knowledge-Based Intelligent Information and Engineering Systems: 9th International Conference, KES 2005*. pp 547-553 (Melbourne, Australia, 2005).

[94] Magnani, L. Creating chances through cognitive niche construction. The role of affordances. In: Apolloni, B.; Howlett, R. J.; Jain, L. (eds) *Lecture Notes In Computer Science: Knowledge-Based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007*. pp 917–925 (Springer, Berlin, 2007).

[95] Manikonda, P. K.; Rajendra, P.; Devendranath, D.; Gunasekaran, B.; Channakeshava; Aradhya, R. S.; Sashidhar, R. B.; Subramanyam, C. Influence of extremely low frequency magnetic fields on Ca2+ signaling and NMDA receptor functions in rat hippocampus. *Neuroscience Letters* **413(2)**, pp 145-149 (2007).

[96] Matson, J. L., LoVullo S. V. Trends and topics in autism spectrum disorders research. *Research in Autism Spectrum Disorders* **3(1)**, pp 252-257 (2009).

[97] Mattson, M. P. NF-kappaB in the survival and plasticity of neurons. *Neurochemical Research* **30(6-7)**, pp 883-893 (2005).

[98] McCain, K. W. Longitudinal author cocitation mapping: The changing structure of macroeconomics. *Journal of the American Society for Information Science* **35**, pp 351-359 (1984).

[99] Mednick, S. A. The associative basis of the creative process. *Psychological Review* **69(3)**, pp 220-232 (1962).

[100] Miller, G. A. WordNet: a lexical database for English. *Communications of the ACM* **38(11)**, pp 39-41 (1995).

[101] Ming, X., Stein, T. P., Brimacombe, M., Johnson, W. G., Lambert, G. H., Wagner, G. C. Increased excretion of a lipid peroxidation biomarker in autism. *Prostaglandins, Leukotrienes, and Essential Fatty Acids* **73(5)**, pp 379-384 (2005).

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

[103] Moore, D. S.; McCabe, G. P. Introduction to the Practice of Statistics, 3rd ed. (W. H. Freeman, New York, 1999).

[104] Morreale de Escobar G, Obregon MJ, Escobar del Rey F. Role of thyroid hormone during early brain development. *European Journal of Endocrinology* Suppl 3:U25-U37 (2004).

[105] Mouridsen, S. E.; Rich, B.; Isager, T.; Nedergaard, N. J. Autoimmune diseases in parents of children with infantile autism: a case-control study. *Developmental Medicine & Child Neurology* **49(6)**, pp 429-32 (2007).

[106] National Institutes of Health Clinical Center. Treatment of childhood regressive autism with Minocycline: An anti-inflammatory agent active within the CNS. Study Start Date: November 2006. http://clinicalstudies.info.nih.gov/detail/A_2007-M-0024.html (accessed October 2007)

[107] Nelson, S. J.; Johnston, D.; Humphreys, B. L. Relationships in Medical Subject Headings. In: Bean, C. A.; Green, R. (eds) *Relationships in the organization of knowledge*. pp 171-184 (Kluwer Academic Publishers, New York, 2001).

[108] Nenadic, G.; Spasic, I.; Ananiadou, S. Terminology-driven mining of biomedical literature. *Bioinformatics* **19(8)**, pp 938-943 (2003).

[109] Ohsawa, Y. Chance discovery: the current states of art. *Chance Discoveries in Real World Decision Making* **30**, pp 3-20 (2006).

[110] Omura, Y. Asbestos as a possible major cause of malignant lung tumors (including small cell carcinoma, adenocarcinoma & mesothelioma), brain tumors (i.e. astrocytoma & glioblastoma multiforme), many other malignant tumors, intractable pain including fibromyalgia, & some cardio-vascular pathology: safe & effective methods of reducing asbestos from normal & pathological areas. *Acupuncture & electro-therapeutics research* **31(1-2)**, pp 61-125 (2006).

[111] Persico, A. M.; Bourgeron, T. Searching for ways out of autism maze: genetic, epigenetic and environmental clues. *Trends in Neurosciences* **29(7)**, pp 349-358 2006.

[112] Petrič, I.; Urbančič, T.; Cestnik, B. Comparison of ontologies built on titles, abstracts and entire texts of articles. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O.; Musek, J.; Osredkar, M. J.; Kononenko, I.; Novak Škarja, B. (eds) I*S-2006. Proceedings of the 9th International multi-conference Information Society*. pp 227-230 (Ljubljana, Slovenia, 2006).

[113] Petrič, I.; Urbančič, T.; Cestnik, B. Discovering hidden knowledge from biomedical literature. *Informatica* **31(1)**, pp 15-20 (2007).

[114] Petrič, I.; Urbančič, T.; Cestnik, B. Literature mining: potential for gaining hidden knowledge from biomedical articles. In: Bohanec, M.; Gams, M.; Rajkovič, V.; Urbančič, T.; Bernik, M.; Mladenić, D.; Grobelnik, M.; Heričko, M.; Kordeš, U.; Markič, O.; Musek, J.; Osredkar, M. J.; Kononenko, I.; Novak Škarja, B. (eds) I*S-2006. Proceedings of the 9th International multi-conference Information Society*. pp 52-55 (Ljubljana, Slovenia, 2006).

[115] Petrič, I.; Urbančič, T.; Cestnik, B.; Macedoni-Lukšič, M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics* **42(2)**, pp 219-227 (2009).

[116] PubMed. Overview. http://www.ncbi.nlm.nih.gov/ (accessed September 2008)

[117] PubMed Central. PMC Overview. http://www.pubmedcentral.nih.gov/about/intro.html (accessed August 2006)

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

[119] Raychaudhuri, S.; Schütze, H.; Altman, R. B. Using text analysis to identify functionally coherent gene groups. *Genome Research* **12(10)**, pp 1582-1590 (2002).

[120] Rindflesch, T. C.; Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* **36(6)**, pp 462-477 (2003).

[121] Roesler, R.; Henriques, J. A. Schwartsmann G. Gastrin-releasing peptide receptor as a molecular target for psychiatric and neurological disorders. *CNS & Neurological Disorders - Drug Targets* **5(2)**, pp 197-204 (2006).

[122] Román, G. C. Autism: transient in utero hypothyroxinemia related to maternal flavonoid ingestion

during pregnancy and to other environmental antithyroid agents. *Journal of the Neurological Sciences* **262(1-2)**, pp 15-26 (2007).

[123] Runyan, J. D.; Moore, A. N.; Dash, P.K. A role for prefrontal calcium-sensitive protein phosphatase and kinase activities in working memory. *Learning & Memory* **12(2)**, pp 103-110 (2005).

[124] Rusnak, F.; Mertz, P.. Calcineurin: form and function. *Physiological Reviews* **80(4)**, pp 1483-1521 (2000).

[125] Salton, G.; Buckley, C. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* **24(5)**, pp 513-523 (1988).

[126] Sayers, E.; Wheeler, D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils) In: *U.S. National Library of Medicine. NCBI Short Courses* (2004).

[127] Schönhofen, P.; Benczúr, A. A. Exploiting Extremely Rare Features in Text Categorization. In: F¨urnkranz, J.; Scheffer, T.; Spiliopoulou M. (eds) *Lecture Notes in Computer Science. Machine Learning: ECML 2006*. pp 759–766 (Springer, Berlin, Heidelberg, 2006).

[128] Schwartz, A. S.; Hearst, M. A. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing*, pp 451-462 (2003).

[129] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys* **34(1)**, pp 1-47 (2002).

[130] Segura-Bedmar, I.; Martínez, P.; Segura-Bedmar, M. Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discovery Today* **13(17-18)**, pp 816-823 (2008).

[131] Sen, R.; Baltimore, D. Multiple nuclear factors interact with the immunoglobulin enhancer sequences. *Cell* **46(5)**, pp 705-716 (1986).

[132] Shih, T. C.; Hsieh, S. Y.; Hsieh, Y. Y.; Chen, T. C.; Yeh, C. Y.; Lin, C. J.; Lin, D. Y.; Chiu, C. T. Aberrant activation of nuclear factor of activated T cell 2 in lamina propria mononuclear cells in ulcerative colitis. *World Journal of Gastroenterology* **14(11)**, pp 1759-1767 (2008).

[133] Shortliffe, E. H. The adolescence of AI in medicine: will the field come of age in the '90s? *Artificial Intelligence in Medicine* **5(2)**, pp 93-106 (1993).

[134] Singhal, A.; Jajodia, S. Data warehousing and data mining techniques for intrusion detection systems. *Distributed and Parallel Databases* **20(2)**, pp 149-166 (2006).

[135] Sinha, A. K.; Pickard, M. R.; Hubank, M. J.; Ruiz de Elvira, M. C.; Hadjzadeh, M.; Attree, E. A.; Davey, M. J.; Rose, F. D.; Ekins, R. P. Maternal hypothyroxinemia and brain development: II. Biochemical, metabolic and behavioural correlates. *Acta Medica Austriaca* **19(Suppl 1)**, pp 49-54 (1992).

[136] Sivagnanasundaram, S.; Fletcher, D.; Hubank, M.; Illingworth, E.; Skuse, D.; Scambler, P. Differential gene expression in the hippocampus of the Df1/+ mice: A model for 22q11.2 deletion syndrome and schizophrenia. *Brain Research* **1139**, pp 48-59 (2007).

[137] Smalheiser, N. R.; Swanson, D. R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine* **57(3)**, pp 149-153 (1998).

[138] Srinivasan, P.; Libbus, B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* **20(Suppl 1)**, pp I290-296 (2004).

[139] Srinivasan, P.; Libbus, B.; Sehgal, A. K. Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases. In: *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*. pp 33-40 (Boston, MA, 2004).

[140] Steele, S. D.; Minshew, N. J.; Luna, B.; Sweeney J. A. Spatial working memory deficits in autism. *Journal of Autism and Developmental Disorders* **37(4)**, pp 605-612 (2007).

[141] Steinbach, W. J.; Reedy, J. L.; Cramer, R. A. Jr; Perfect, J. R.; Heitman, J. Harnessing calcineurin as a novel anti-infective agent against invasive fungal infections. *Nature Reviews Microbiology* **5(6)**, pp 418-430 (2007).

[142] Suzuki, E.; Kodratoff, Y. Discovery of surprising exception rules based on intensity of implication. In: Zytkow, J. M.; Quafafou, M. (eds) *Lecture Notes in Computer Science. Principles of Data Mining and Knowledge Discovery. Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*. pp 10-18 (Springer, Berlin, 1998).

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

[144] Swanson, D. R. Undiscovered public knowledge. *Library Quarterly* **56(2)**, pp 103-118 (1986).

[145] Swanson, D.R. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine* **31**, pp 526–557 (1988).

[146] Swanson, D. R.; Smalheiser, N. R.; Torvik, V. I. Ranking indirect connections in literature-based discovery: The role of Medical Subject Headings (MeSH). *Journal of the American Society for Information Science and Technology* **57(11)**, pp 1427-1439 (2006).

[147] Štěpánková, O.; Engová, D. Professional competence and computer literacy in e-age, focus on healthcare. *Methods of Information in Medicine* **45(3)**, pp 300-304 (2006).

[148] Tanabe, L.; Wilbur, W. J. Tagging gene and protein names in biomedical text. *Bioinformatics* **18(8)**, pp 1124-1132 (2002).

[149] Thelwall, M. *Link Analysis: An Information Science Approach* (Elsevier Academic Press, Amsterdam, 2004).

[150] Thibeault, I.; Laflamme, N.; Rivest, S. Regulation of the gene encoding the monocyte chemoattractant protein 1 (MCP-1) in the mouse and rat brain in response to circulating LPS and proinflammatory cytokines. *The Journal of Comparative Neurology* **434(4)**, pp 461-477 (2001).

[151] Thomas, P. G.; Carter, M. R.; Da'dara, A. A.; DeSimone, T. M.; Harn, D. A. A helminth glycan induces APC maturation via alternative NF-kappa B activation independent of I kappa B alpha degradation. *The Journal of Immunology* **175(4)**, pp 2082-2090 (2005).

[152] Thornton, I. M. Out of time: a possible link between mirror neurons, autism and electromagnetic radiation. *Medical Hypotheses* **67(2)**, pp 378-382 (2006).

[153] U.S. National Library of Medicine. 2000 Cumulated Index Medicus: The End of an Era. *NLM Technical Bulletin* **321**, e3 (2001).

[154] U.S. National Library of Medicine. Fact SheetUMLS® Metathesaurus®. Published: 28 March 2006. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html (accessed September 2007).

[155] U.S. National Library of Medicine. Fact SheetUnified Medical Language System®. Published: 23 March 2006. http://www.nlm.nih.gov/pubs/factsheets/umls.html (accessed October 2007).

[156] U.S. National Library of Medicine. MEDLINE® Citation Counts by Year of Publication. Published: 27 May 2003. http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html (accessed September 2008).

[157] U.S. National Library of Medicine. MEDLINE® PubMed® XML Element Descriptions and their Attributes. Published: 12 December 2005. http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html (accessed February 2007).

[158] U.S. National Library of Medicine. NLM® uses XML for MEDLINE® Data. Published: 31 August 2000. http://www.nlm.nih.gov/news/medlinedata.html (accessed February 2007).

[159] United Nations. Resolution adopted by the General Assembly. A/RES/62/139. World Autism Awareness Day. 18 December 2007 http://www.un.org/observances/days.shtml (accessed September 2008).

[160] Urbančič, T.; Petrič, I.; Cestnik, B.; Macedoni-Lukšič, M. Literature mining: towards better understanding of autism. In: Bellazzi, R.; Abu-Hanna, A.; Hunter, J. (eds) *AIME 2007. Proceedings of the 11th Conference on Artificial Intelligence in Medicine in Europe*. pp 217-226 (Amsterdam, The Netherlands, 2007).

[161] Vargas, D. L.; Nascimbene, C.; Krishnan, C.; Zimmerman, A. W.; Pardo, C. A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Annals of Neurology* **57(1)**, pp 67-81 (2005).

[162] Vorstman, J. A.; Morcus, M. E.; Duijff, S. N.; Klaassen, P. W.; Heineman-de Boer, J. A.; Beemer, F. A.; Swaab, H.; Kahn, R. S.; van Engeland, H. The 22q11.2 deletion in children: high rate of autistic disorders and early onset of psychotic symptoms. *Journal of the American Academy of Child & Adolescent Psychiatry* **45(9)**, pp 1104-1113 (2006).

[163] Vosgerau, H. Migraine therapy with magnesium glutamate. *Therapie der Gegenwart* **112(4)**, pp 640 (1973).

[164] Wang, J. H.; Desai, R. Modulator binding protein. Bovine brain protein exhibiting the Ca2+-

dependent association with the protein modulator of cyclic nucleotide phosphodiesterase. *Journal of Biological Chemistry* **252(12)**, pp 4175-4184 (1977).

[165] Weeber, M. Drug discovery as an example of literature-based discovery. In: Džeroski, S.; Todorovski, L. (eds) *Lecture Notes in Computer Science. Computational Discovery of Scientific Knowledge*. pp 290-306 (Springer, Berlin, Heidelberg, 2007).

[166] Weeber, M.; Vos, R.; Klein, H.; de Jong-van den Berg, L. T. W. Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology* **52(7)**, pp 548-557 (2001).

[167] Weeber, M.; Vos, R.; Klein, H.; de Jong-van den Berg, L. T. W.; Aronson, A. R.; Molema, G. 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 (2003).

[168] Weiss, G. M. Mining with rare cases. In: Maimon, O.; Rokach, L. (eds) *Data Mining and Knowledge Discovery Handbook* . pp 765-776 (2005).

[169] Wiley, S.; Choo, D.; Meinzen-Derr, J.; Hilbert, L.; Greinwald, J. GJB2 mutations and additional disabilities in a pediatric cochlear implant population. *International Journal of Pediatric Otorhinolaryngology* **70(3)**, pp 493-500 (2006).

[170] Winder, D. G.; Sweatt, J. D. Roles of serine/threonine phosphatases in hippocampal synaptic plasticity. *Nature Reviews Neuroscience* **2(7)**, pp 461-474 (2001).

[171] Winter, W. E.; Schatz, D. Prevention strategies for type 1 diabetes mellitus: current status and future directions. *BioDrugs* **17(1)**, pp 39-64 (2003).

[172] Wood, J. G.; Wallace, R. W.; Whitaker, J. N.; Cheung, W. Y. Immunocytochemical localization of calmodulin and a heat-labile calmodulin-binding protein (CaM-BP80) in basal ganglia of mouse brain. *The Journal of Cell Biology* **84(1)**, pp 66-76 (1980).

[173] Wu, Z.; Tawfik, A. Y. Towards a change-based chance discovery. In: Chen, C. S.; Filipe, J.; Seruca, I.; Cordeiro, J. (eds) *Enterprise Information Systems VII*. pp 131–138 (Springer, Dordrecht, The Netherlands, 2006).

[174] Yamauchi, M.; Tamaki, S.; Tomoda, K.; Yoshikawa, M.; Fukuoka, A.; Makinodan, K.; Koyama, N.; Suzuk,i T.; Kimura, H. Evidence for activation of nuclear factor kappaB in obstructive sleep apnea. *Sleep Breath* **10(4)**, pp 189-193. 2006

[175] Yerys, B. E.; Hepburn, S. L.; Pennington, B. F.; Rogers, S. J. Executive function in preschoolers with autism: evidence consistent with a secondary deficit. *Journal of Autism and Developmental Disorders* **37(6)**, pp 1068-1079 (2007).

[176] Yetisgen-Yildiz, M.; Pratt, W. Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics* **39(6)**, pp 600-611 (2006).

[177] Yoo, H. J.; Cho, I. H.; Park, M.; Cho, E.; Cho, S. C.; Kim, B. N.; Kim, J. W.; Kim, S. A. Association between PTGS2 polymorphism and autism spectrum disorders in Korean trios. *Neuroscience Research* **62(1)**, pp 66-69 (2008).

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

[179] Zou, J.; Crews, F. CREB and NF-kappaB Transcription Factors Regulate Sensitivity to Excitotoxic and Oxidative Stress Induced Neuronal Cell Death. *Cellular and Molecular Neurobiology* **26(4-6)**, pp 383-403 (2006).

[180] Zupan, B.; Bohanec, M.; Bratko, I.; Cestnik, B. A Dataset Decomposition Approach to Data Mining and Machine Discovery. In: Heckerman, D. (ed.) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. pp 299-302 (Newport Beach, California, 1997).

Index of Figures

Figure 1: *Text mining process*. The sequence of steps is modelled in conformity with a definition of knowledge discovery in databases (KDD) process as originally proposed by Fayyad and colleagues. 7

Figure 2: *Citation counts of MEDLINE publications*. The numbers of documents cited in MEDLINE in the period from 1948 to 2007 (Source: U.S. National Library of Medicine, April 2008). 8

Figure 3: *Terror news network analysed by Pajek (a popular link analysis tool) based on news reports following the September 11 attack on the United States (Batagelj and Mrvar, 2003)*. In the example network, the thickness of an edge represents the frequency of co-appearance of two words that are linked by the given edge (Source: http://vlado.fmf.uni-lj.si/pub/networks/pajek/pics/examples.htm, January, 2008). 11

Figure 4: *Closed (left model) versus open (right model) discovery process as defined by Weeber et al. (Weeber et al., 2001)*. 14

Figure 5: *Open discovery process as applied by Srinivasan and colleagues*. 15

Figure 6: *Open discovery process in LitLinker*. 16

Figure 7: *Growth of autism publications in comparison with the growth of all citations in MEDLINE*. The numbers of autism citations and the numbers of all documents cited in MEDLINE in the period from 1948 to 2007 are presented with two different sets of values for the Y-axis scale. 18

Figure 8: *Screenshot of PubMed that provides access to the articles indexed for MEDLINE*. PubMed is available via the Entrez retrieval system, developed by the U.S. National Center for Biotechnology Information at the National Library of Medicine, located at the U.S. National Institutes of Health. 22

Figure 9: *The sample record from the MEDLINE database in the XML format*. The displayed fields of the sample record are part of our article from the Journal of Biomedical Informatics, viewed as MEDLINE citation in XML format (Source: U.S. National Library of Medicine, September 2008). 24

Figure 10: *MeSH descriptor data for autistic disorder*. (Source: U.S. National Library of Medicine, July 2008). 25

Figure 11: *MeSH tree structure for the Mental Disorders category [F03] when approaching to the Autistic Disorder category [F03.550.325.125]*. (Source: U.S. National Library of Medicine, September 2008). 26

Figure 12: *Autism trend analysis*. Trend of autism research is represented by the ontology, where documents were divided into 5 periods according to the year of their publication. 28

Figure 13: *A two-level autism ontology*. Concepts are renamed according to autism survey literature, based on the keywords suggested by OntoGen. 29

Figure 14: *Screenshot of OntoGen, version 2.0.0.0*. A tool for interactive topic ontology construction (Fortuna et al., 2006). 30

Figure 15: *Top-level autism ontology concepts*. Original concepts descriptions (three for each concept) as suggested by OntoGen are included for easier identification. 32

Figure 16: *Comparison between the distributions of documents when they were divided into 8 sub-concepts* 34

Figure 17: *Comparison between the distributions of documents when they were divided into 5 sub-concepts* 36

Figure 18: *Comparison between the distributions of documents belonging to the ontology concepts of abstracts and bodies of texts when documents were divided into 8 sub-concepts* 37

Figure 19: *Comparison between the distributions of documents belonging to the ontology concepts of abstracts and bodies of texts when documents were divided into 5 sub-concepts*. 38

Figure 20: *Combined open and closed discovery process in the RaJoLink method*. The upper half of the figure corresponds to the open discovery (identifying rare terms *r* and finding a joint term *a*) and the lower half to the closed discovery (searching for linking terms *b*). 41

Figure 21: *Similarity graph representing instances of literature A and instances of literature C according to their content similarity*. The distinctive outliers are positioned far enough away from the most typical representatives of the two heretofore unrelated literatures. 44

Figure 22: *Flow chart showing the procedures of the RaJoLink method*. 45

Figure 23: *The number of terms according to their total frequencies in the set of abstracts that were available from 11,781 articles on autism published in MEDLINE until the end of year 2007*. The total frequency of a term (X-axis) represents the number of abstracts that contain that particular term. The Y-axis value reveals how many terms appear with a given total frequency in the set of records. 46

Figure 24: *The number of rare terms according to their internal frequencies observed in the set of abstracts that were available from 11,781 articles on autism published in MEDLINE until the end of year 2007*. The internal frequency of a term (X-axis) is determined by the number of times the term appears within a single abstract. 47

Figure 25: *A screenshot of the RaJoLink system showing second level MeSH categories (V01, V02, V03, VO4) and the general category of terms (V05) within their top-level category (V).* 48

Figure 26: *OntoGen's similarity graph of a set of autism and calcineurin articles' abstracts*. Two main article topics (*AUTISM* and *CALCINEURIN*) are listed on the left side of the window. As the autism topic is selected, the list of abstracts, which are in the relationship with this selected topic, is presented in the central part of the OntoGen's window. The distinctive calcineurin article (*CN 437*) is visualized among the autism context documents. 50

Figure 27: *Schema of the RaJoLink method*. 51

Figure 28: *A screenshot of RaJoLink where only rare terms from autism documents are displayed (parameter Frequency =1)*. MeSH codes for each term are listed on the right hand side. The terms chromogranin and cofilin are selected (checked) for further analysis. 54

Figure 29: *A screenshot of RaJoLink showing part of results for candidate joint terms*. The term calcineurin is selected for further analysis. 55

Figure 30: *A screenshot of the proposed user interface showing the visualization of results for candidate linking terms*. Each candidate linking term would be highlighted in pairs of articles from literature *A* and from literature *C*. 57

Figure 31: *The Ra step*. Overview of rare terms detection within the RaJoLink's literature-based knowledge discovery approach. 60

Figure 32: *The calcineurin signalling pathway in T-cells.* Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Microbiology (Steinbach et al., 2007), copyright (2007). 62

Figure 33: *A two-level calcineurin ontology*. Concepts are named with the keywords suggested by OntoGen. 63

Figure 34: *Venn diagram of arguments (b_i) found as connection between the scientific literature on autism (C) and the scientific literature on calcineurin (A).* 63

Figure 35: *Experimental results obtained on autism+fragile_X domain*. 65

Figure 36: *A two-level ontology that captures the view of NF-kappaB domain.* Concepts are named with the keywords suggested by OntoGen. 66

Figure 37: *OntoGen's similarity graph of a combined set of autism and NF-kappaB articles' abstracts*. The central part shows documents that belong to the currently selected NF-kappaB main topic. The article on autism is marked as an exception within the selected topic and positioned among the NF-kappaB context documents. 67

Figure 38: *The NF-kappaB activation pathway.* NF-kappaB is mainly sequestered in the cytoplasm bound to inhibitory IkappaB-alpha proteins. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology (Chen and Greene, 2004), copyright (2004). 69

Figure 39: *The results of the RaJoLink method applied to the domain of migraine.* Besides magnesium, another 3 discoveries were obtained with the RaJoLink system, namely the joint terms: interferon, interleukin and tnf..... 76

Figure 40: *The document frequencies of rare terms that appeared together with the term magnesium in the titles of MEDLINE documents published before 1988.* The rare terms (e.g. the term *seizures*) that appeared in more than five titles of articles together with the term *magnesium* are marked with their series name (i.e. the MeSH category) and the point name (i.e. the observed rare term). 78

Figure 41: *A screenshot of RaJoLink showing filtered results in step Ra for the migraine-magnesium experiment*. The selected rare terms from the literature about migraine are chosen for further analysis. 79

Figure 42: *A screenshot of RaJoLink showing filtered results for candidate joint terms*. The terms magnesium, interferon and interleukin appear on the top 30 list of candidate joint terms..... 80

Index of Tables

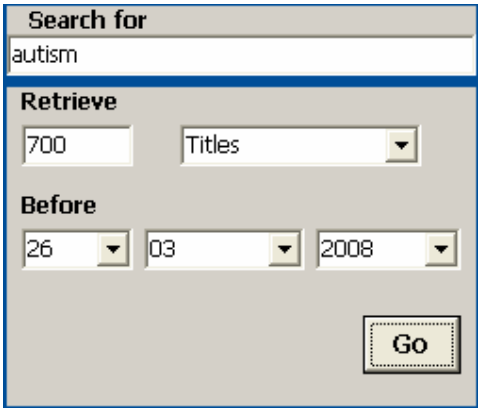
Table 1: *Eight concepts of autism ontology generated from 214 titles.* 33
Table 2: *Eight concepts of autism ontology generated from 214 abstracts.* 33
Table 3: *Eight concepts of autism ontology generated from 214 bodies of texts.* 33
Table 4: *Five concepts of autism ontology generated from 214 titles.* 35
Table 5: *Five concepts of autism ontology generated from 214 abstracts*..... 35
Table 6: *Five concepts of autism ontology generated from 214 bodies of texts.* 35
Table 7: *Comparison of individual keywords extracted from concepts names of autism ontologies when documents were divided into 5 sub-concepts.* 39
Table 8: *Comparison of individual keywords extracted from concepts names of autism ontologies when documents were divided into 8 sub-concepts.* 39
Table 9: *Hypotheses for autism and calcineurin relationship*. Pairs of MEDLINE articles that connect some autism findings on the one hand to the specific calcineurin observations on the other hand..... 64
Table 10: *Hypotheses for autism and NF-kappaB relationship*. Pairs of MEDLINE articles that connect specific autism findings on the one hand to the NF-kappaB observations on the other hand..... 68

Appendix: RaJoLink – User Manual

The current version of the RaJoLink application with detailed installation instructions can be obtained by sending a request by email to ingrid.petric@ung.si.

When the installation is done, start the RaJoLink application. Each research must begin in the main search window with the phase "Rare". You can either locate literature (e.g., bodies of articles) by browsing manually for a specific text file or you can automatically retrieve titles or abstracts of the MEDLINE articles using a simple word search. In any case check the parts of articles you want to analyze.

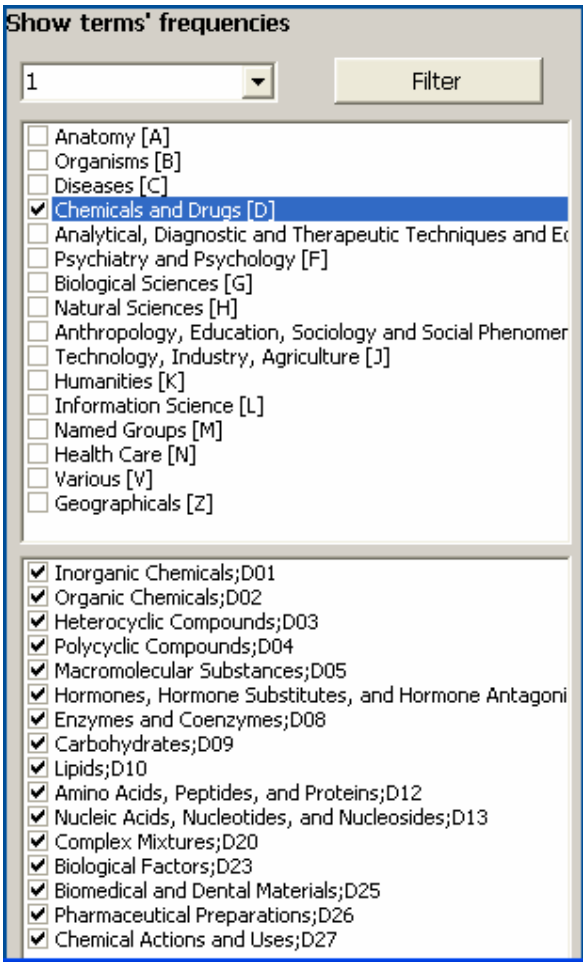
Start the search for rare terms. In the search window, type a search query in the *Search for* text box. You can enter your search query in a natural language format (e.g., autistic disorder) or as a Boolean search string comprised of terms connected by Boolean operators (i.e., AND, NOT, OR). For best results, enclose phrases in quotation marks (e.g., “autistic disorder”). You can additionally limit your search results to articles published in a date range up to a desired date or by limiting the number of articles – in this case RaJoLink displays the most recent MEDLINE publications first. Click the *Go* button to perform the query.



RaJoLink searches for documents matching your search query, and displays a list of matching texts and a statistics of terms in the form of a result list.

Filter the results. RaJoLink offers filtering according to Medical Subject Headings (MeSH) that enable you to limit your search results to a particular medical subject and to maximal number of terms frequency. Focusing on such words can narrow down the search space and, thus, speed-up and improve the inference process.

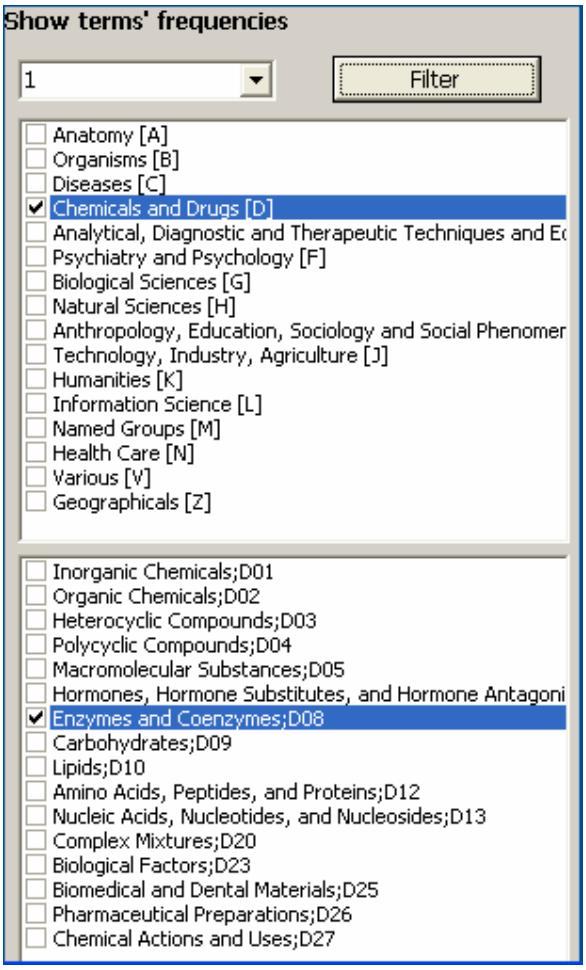
We use the second-level categories from the 2008 MeSH tree structure (i.e., Behaviour and Behaviour Mechanisms- F01, Psychological Phenomena and Processes - F02, Mental Disorders - F03, Behavioural Disciplines and Activities - F04) to classify terms from the input text collection. Each of the second-level categories belongs to one of the top-level categories in the MeSH hierarchy. In the example below, the Chemicals and Drugs [D] main heading is selected.



Click the *Filter* button to perform filtering. The number of resulting terms substantially decreases according to the performed filtering.

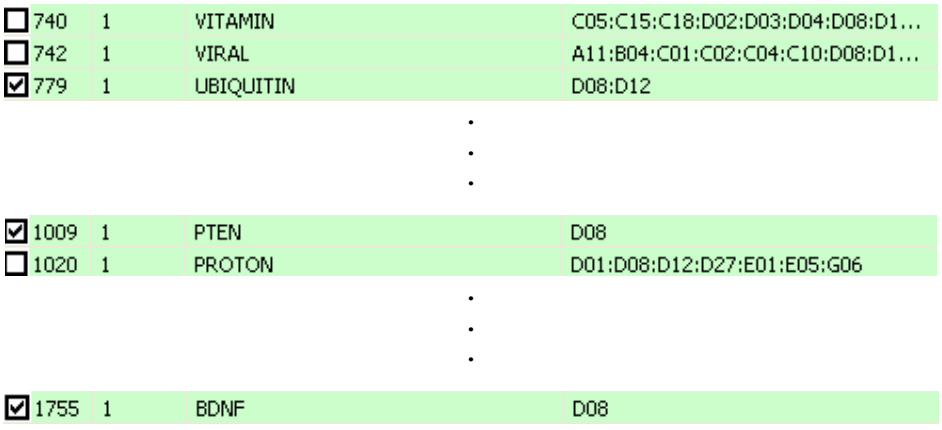
☐	717	1	ZERO	D12:G03:H01
☐	721	1	XI	C15:D05:D08:D12
☐	729	1	WEIGHT	C23:D05:D09:D12:D13:D27:E01:E05:F02:H01:I03:M01
☐	740	1	VITAMIN	C05:C15:C18:D02:D03:D04:D08:D10:D12:D27:G06
☐	742	1	VIRAL	A11:B04:C01:C02:C04:C10:D08:D12:D13:D20:D23:E01:E05:G04:G05:G14
☐	756	1	VASOPRESSIN	D06:D12
☐	762	1	VACCINES	D03:D12:D20
☐	780	1	UBIQUITIN	D08:D12

For instance, suppose that your main interest was in enzymes and coenzymes that could influence the phenomenon under research. In this case you would choose only the MeSH category D08, to which enzymes and coenzymes are designated as illustrated also in figure below. Click the *Filter* button again to perform filtering.

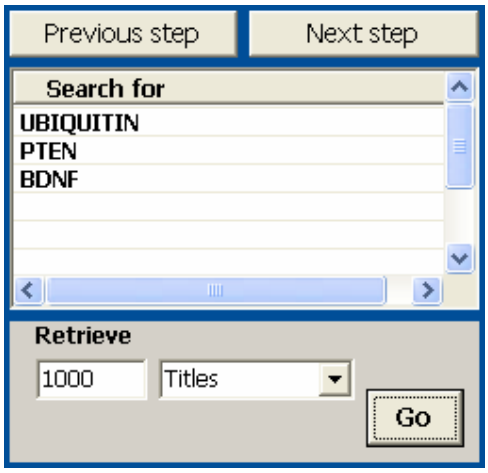


By such filtering the number of resulting terms decreases to a smaller number of those terms that in the example case refer to enzymes and coenzymes. To expand the original query or perform another search, define different query parameters in the *Search for* text box and/or type a new term and click the *Go* button.

Select the interesting rare terms. Before proceeding to the next step you have to choose at least two interesting rare terms (e.g., "UBIQUITIN", "PTEN" and "BDNF" in the above example) that will be analyzed in the "Joint" step. To go to the next step of the method click the *Next step* button.



Start the query for joint terms. In the search window appear the previously chosen rare terms in the *Search for* text box. You can manually add more terms or change (rename, delete) the proposed ones by selecting a row and clicking on it.



For each rare term, the system displays the relating titles or abstracts in the left bottom window and the terms statistics in the main window. Like in previous step, the total frequencies of terms are computed for each set of records about the selected rare terms. The list of resulting terms is compound from terms, their total frequencies in sets of records for each of the selected rare terms and their MeSH codes. All the frequencies are summed in the column *Sum of frequencies*. This number is added to the list of joint terms in order to facilitate the validation of terms regarding their possibility to be selected as a promising joint term.

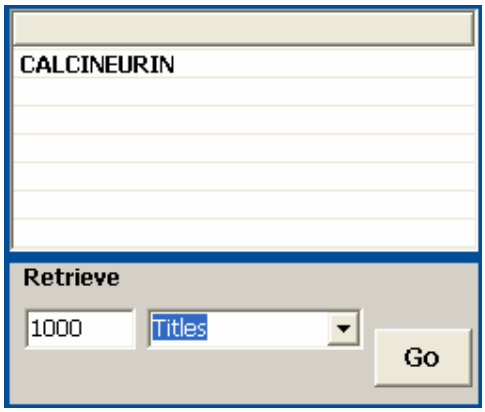
The system displays only those terms that were not found in the corpus of records about the starting term (autism in our case), and therefore pass the criterion of possible hypothesis generation. The records can be sorted by clicking on the column title.

ID	Term	Frequencies	Sum of frequencies	MeSH codes
☐ 0	XENOPUS	3:2,2,4	8	B01:D12
☐ 1	VIVO	3:8,8,15	31	V05
☐ 2	VITRO	3:10,9,20	39	E02:E05:V03
☐ 3	VIABILITY	3:1,3,3	7	D12:G04:G07
☐ 4	VESICLE	3:3,1,1	5	A05:A08:A11:D12
☐ 5	VECTOR	3:1,4,9	14	E01:G03:G14
☐ 6	VASCULAR	3:4,11,2	17	A02:A04:A07:A08:A11:C04:C06...

Filter the results and select the interesting joint term/s. When you are satisfied with search results continue with the next step of the method by clicking the *Next step* button.

ID	Term	Frequencies	Sum of frequencies	MeSH codes
☑ 273	CALCINEURIN	3:2,1,2	5	D08

Start the query for linking terms. In the *Search for* text box of the search window appear the previously chosen joint term/s. You can manually change the proposed term/s by selecting a row and clicking on it to insert a new term or change the existing one. After clicking on the *Go* button, the system automatically retrieves articles on the selected joint term/s from MEDLINE and considers their analysis together with the analysis of the starting literature.



Filter the results and display the significant linking terms (e.g., interleukin in the autism example case).

	Term	Frequencies	Sum of frequencies	MeSH codes
342	INTERLEUKIN	2:3,1	4	D08:D12
347	INTEGRIN	2:2,1	3	D12
349	INSULIN	2:6,1	7	A03:C10:C18:D06:D08:D12:D27:...
354	INHIBITORY	2:6,2	8	D05:D06:D08:D12:E05:G07
355	INHIBITION	2:33,1	34	D12:E01:E02:F01:G04:G07:G08:...
377	IMMUNE	2:7,5	12	A12:A15:B04:C02:C10:C13:C20:...
378	IMMEDIATE	2:1,2	3	C13:C20:D12:E06:F02:G14
382	IGG	2:1,1	2	C15:D12
387	HUMAN	2:38,7	45	A01:A03:A08:A11:A12:B01:B04:...
388	HOMOLOG	2:1,1	2	D08:D12:G06
391	HIGH	2:10,33	43	A08:B03:C04:C09:C10:C14:C16:...
395	GUIDANCE	2:3,2	5	D12:E06:F02:N02
396	GROWTH	2:22,8	30	A02:A06:A08:C04:C05:C13:C23:...
398	GLUTAMATE	2:4,3	7	D08:D12:D27
404	GENERAL	2:1,1	2	C01:C10:D12:D27:E03:F04:G02:...
408	GABA	2:2,3	5	D02:D08:D12:D27

At any time, you may return to the previous steps of the method by clicking the *Previous step* button and enter a different search. To display the records that satisfy new search criteria, click on the *Go* button again.