Bisociative Knowledge Discovery by Literature Outlier Detection

Ingrid Petrič[1], Bojan Cestnik[2,3], Nada Lavrač[3,1], and Tanja Urbančič[1,3]

[1] University of Nova Gorica, Nova Gorica, Slovenia
{ingrid.petric,tanja.urbancic}@ung.si
[2] Temida, d.o.o., Ljubljana, Slovenia
bojan.cestnik@temida.si
[3] Jožef Stefan Institute, Ljubljana, Slovenia
nada.lavrac@ijs.si

Abstract. The aim of this chapter is to present the role of outliers in literature-based knowledge discovery that can be used to explore potential bisociative links between different domains of expertise. The proposed approach upgrades the RaJoLink method which provides a novel framework for effectively guiding the knowledge discovery from literature, based on the principle of rare terms from scientific articles. This chapter shows that outlier documents can be successfully used as means of detecting bridging terms that connect documents of two different literature sources. This linking process, known also as closed discovery, is incorporated as one of the steps of the RaJoLink methodology, and is performed by using publicly available topic ontology construction tool OntoGen. We chose scientific articles about autism as the application example with which we demonstrated the proposed approach.

Keywords: outliers, bisociations, literature mining, knowledge discovery.

1 Introduction

In statistics, an outlier is described as an observation that is numerically distant from the rest of the data, or more formally, it is an observation that falls outside the overall pattern of a distribution [1]. While in many data sets outliers may be due to data measurement errors (therefore it would be best to discard them from the data), there are also several examples where outliers actually led to important discoveries of intriguing information.

In this chapter we explore the potential of outliers for guiding bisociative knowledge discovery from literature. We present an approach to outliers-based knowledge discovery from text documents that can be used to explore implicit relationships across different domains of expertise, indicating interesting cross-domain connections, called *bisociations* [2], [3]. The development of this approach was conducted in three phases that were described in a comprehensive report [4]. The approach upgrades the RaJoLink method [5] for knowledge discovery from literature,

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 313–324, 2012.

where the hypotheses generation phase is based on the principle of rare terms from scientific articles, with the notion of bisociation.

The motivation for work has grounds in the associationist creativity theory [8]. Mednick [8] defines creative thinking as the faculty of generating new combinations of distant associative elements (e.g. words). He explicates how thinking of concepts that are not strictly related to the elements under research inspires unforeseen useful connections between elements. In this manner, bisociations considerably improve the knowledge discovery process. This chapter pays special attention to the category of context-crossing associations, called bisociations [3].

RaJoLink is intended to support experts in their overall process of open knowledge discovery, where hypotheses have to be generated, followed by the closed knowledge discovery process, where hypotheses are tested. It was demonstrated in [5], [6], and [7] that this method can successfully support the user-guided knowledge discovery process.

The RaJoLink methodology has been applied to a challenging medical domain: the set of records for our study was selected from the domain of autism. Autism belongs to a group of pervasive developmental disorders that are portrayed by an early delay and abnormal development of cognitive, communication and social interaction skills of a person [9]. It is a very complex and not yet sufficiently understood domain, where precise causes are still unknown, hence we have chosen it as our experimental testing domain.

This chapter is organized as follows. Section 2 presents the related work in the area of literature mining. Section 3 introduces the literature-based knowledge discovery process and further explores rarity as a principle for guiding the knowledge discovery in the upgraded RaJoLink method. Section 4 presents the RaJoLink approach by focusing on outliers in the closed discovery process. Section 5 illustrates the application of outlier detection to the autism literature. Section 6 provides discussion and conclusions.

2 Related Work in Literature Mining

Novel interesting connections between disparate research findings can be extracted from the published literature. Analysis of implicit associations hidden in scientific literature can guide the hypotheses formulation and lead to the discovery of new knowledge. To support such literature-based discoveries in medical domains, Swanson has designed the *ABC model* approach [10] that investigates whether an agent *A* influences a phenomenon *C* by discovering complementary structures via interconnecting phenomena *B*. Two literatures are complementary if one discusses the relations between *A* and *B*, while a disparate literature investigates the relations between *B* and *C*. If combining these relations suggests a previously unknown meaningful relation between *A* and *C*, this can be viewed as a new piece of knowledge that might contribute to a better understanding of phenomenon *C*.

Weeber and colleagues [11] defined the hypothesis generation approach as an open discovery process and the hypothesis testing as a closed discovery process. In the open discovery process only the phenomenon under investigation (*C*) is given in advance, while the target agent *A* is still to be discovered. In the closed discovery process, both *C* and *A* are known and the goal is to search for bridging phenomena *B* in order to support the validation of the hypothesis about the connection between *A* and *C*. Smalheiser and Swanson [12] developed an online system named ARROWSMITH, which takes as input two sets of titles from disjoint domains *A* and *C* and lists terms *b* that are common to literature *A* and *C*; the resulting terms *b* are used to generate novel scientific hypotheses.[1] As stated by Swanson [13], his major focus in literature-based discovery has been on the closed discovery process, where both *A* and *C* have to be specified in advance.

Several researchers have continued Swanson's line of research. Most of them have made literature-based discoveries in the field of biomedicine. In biomedicine, huge literature databases and well structured knowledge based-systems provide effective supports for literature mining tasks. An on-line literature-based discovery tool called BITOLA has been designed by Hristovski [14]. It uses association rule mining techniques to find implicit relations between biomedical terms. Weeber and colleagues [15] developed Literaby, the concept-based Natural Language Processing tool. The units of analysis that are essential for their approach are UMLS Metathesaurus concepts. The open discovery approach developed by Srinivasan and colleagues [16], on the other hand, relies almost completely on Medical Subject Headings (MeSH). Yetisgen-Yildiz and Pratt [17] proposed a literature-based discovery system called LitLinker. It mines biomedical literature by employing knowledge-based and statistical methods. All the pointed systems use MeSH descriptors [18] as a representation of scientific medical documents, instead of using title, abstract or full-text words. Thus, problems arise since MeSH indexers normally use only the most specific vocabulary to describe the topic discussed in a document [19] and therefore some significant terminology from the documents' content may not be covered. The Swanson's literature-based discovery approach has been extended also by Lindsay and Gordon [20], who used lexical statistics to determine relative frequencies of words and phrases. In their open discovery approach they search for words on the top of the list ranked by these statistics. However, their approach fails when applied to Swanson's first discoveries and extensive analysis has to be based on human knowledge and judgment.

Unlike related work, we put an emphasis on rare terms. Since rare terms are considered to be special terms, not characteristic for a particular domain context, they are more informative than frequent terms. For this reason, rare terms are very likely to be relevant for crossing the boundaries of domains and leading to some interesting observations.

3 The Upgraded RaJoLink Knowledge Discovery Process

The aim of knowledge discovery presented in this chapter is to detect the previously unnoticed concepts (chances) at the intersections of multiple meaningful scenarios. As a consequence, tools for indicating rare events or situations prove to play a significant

[1] Here we use the notations *A*, *B*, and *C* that are written in uppercase letter symbols to represent a set of terms (e.g., literature or collection of titles, abstracts or full texts of documents), while with *a*, *b*, and *c* (lowercase symbols) we represent a single term.

role in the process of research and discovery [21]. From this perspective curious or rare observations of phenomena can provide novel possible opportunities for reasoning [22]. Regarding this, the use of data mining tools is essential to support experts, in choosing meaningful scenarios.

Outliers actually attract a lot of attention in the research world and are becoming increasingly popular in text mining applications as well. Detecting interesting outliers 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 [22].

The rarity principle that we apply in the first (open discovery) step of the RaJoLink literature-based discovery is a fundamental difference from the previously proposed methods and represents a unique contribution of the RaJoLink method. In our earlier work [5], [6], and [7] we presented the idea of extending the Swanson's ABC model to handle the open discovery process with rare terms from the domain literature. For that purpose we employed the Txt2Bow utility from the TextGarden library [23] in order to compute total frequencies of terms in the entire text corpus/corpora.

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. Note that the steps *Ra* and *Jo* implement the open discovery, while the step *Link* corresponds to the closed discovery. The methodological description of the three steps has been provided in our previous publications [5], [6], and [7].

We developed a software tool that implements the RaJoLink method and provides decision support to experts. It can be used to find scientific articles in MEDLINE database [24], to compute statistics about the data, and to analyze them to discover eventually new knowledge. By such exploration, massive amounts of textual data are automatically collected from databases, and text mining methods are employed to generate and test hypotheses. In the 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 the 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 and selected as the candidates for *A*. In order to provide explanation for hypotheses generated in the step *Jo*, our method searches for links between the literature on joint term *a* and the literature on term *c*.

The upgraded RaJoLink methodology for bisociative knowledge discovery consists of the following steps.

- The crucial step in the RaJoLink method is to identify rare elements within scientific literature, i.e., terms that rarely appear in articles about a certain phenomenon.
- Sets of literature about rare terms are then identified and considered together to formulate one or more initial hypotheses in the open discovery process.
- Next, in the closed discovery process, RaJoLink focuses on outlying and their neighbouring documents in the documents' similarity graphs. We construct such graphs with the computational support of a semi-automatic tool for topic ontology construction, called OntoGen [25].

- Outlier documents are then used as a heuristic guidance to speed-up the search for the linking terms (bridging terms, also called b-terms) between different domains of expertise and to alleviate the burden from the expert in the process of hypothesis testing. In this way, the detection of outlier documents represents an upgrade to our previous method that results in significant improvements of the closed discovery process. This step of the upgraded RaJoLink methodology is the focus of research presented in this chapter.

4 Outlier Detection in the RaJoLink Knowledge Discovery Process

This chapter focuses on the steps of the closed discovery process, where two domains of interest *A* and *C* have already been identified prior to starting the knowledge discovery process. The closed discovery process is supported by using the OntoGen tool [25]. One of its features is its capacity of visualizing the similarity between the selected documents of interest. The main novelty of the upgraded RaJoLink methodology is the visualization of outlier documents in the documents' similarity graph (Figure 1) which enables us to find bisociations in the combined set of literatures *A* and *C*. Our argumentation is that outlier documents of two implicitly linked domains can be used to search for relevant linking terms (bridging terms or b-terms) between the two domains. 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 bisociative links is a unique aspect of our method in comparison to the literature-based discovery investigated by other researchers.

When investigating whether disjoint domains *A* and *C* can be connected by domain bridging concepts *B*, we take as input two sets of documents from disjoint domains *A* and *C* and visualize them in the documents' similarity graph. The goal of constructing such graphs is to discover complementary structures that are common to both literatures, *A* and *C* via domain bridging concepts *B*. These domain bridging terms can be found in similarity graphs in those outlying documents of literature *A* and/or literature *C* that are not positioned in the mainstream domain literatures but are relatively distinct from a prototypical document of each domain literature, where a prototypical/average document is, technically speaking, computed as the centroid of the selected domain. Such outlying documents are most frequently observed at the intersection between literatures *A* and *C* as shown in Figure 1.

In the closed discovery process of the RaJoLink method, text documents containing terms *b* that bridge the literature *A* and the literature *C* can be expected to be present in outlier documents. Therefore, in our approach to closed knowledge discovery, outliers are used as heuristic guidance to speed up the search for bridging concepts between different domains of expertise. Having disparate literatures *A* and *C*, both domains are examined by the combined dataset of literatures *A* and *C* in order to assess whether they can be connected by implicit relations. Within the whole corpus of texts consisting of literatures *A* and *C*, which acts as input for step Link (i.e. the closed discovery) of RaJoLink, each text document represents a separate instance/record.

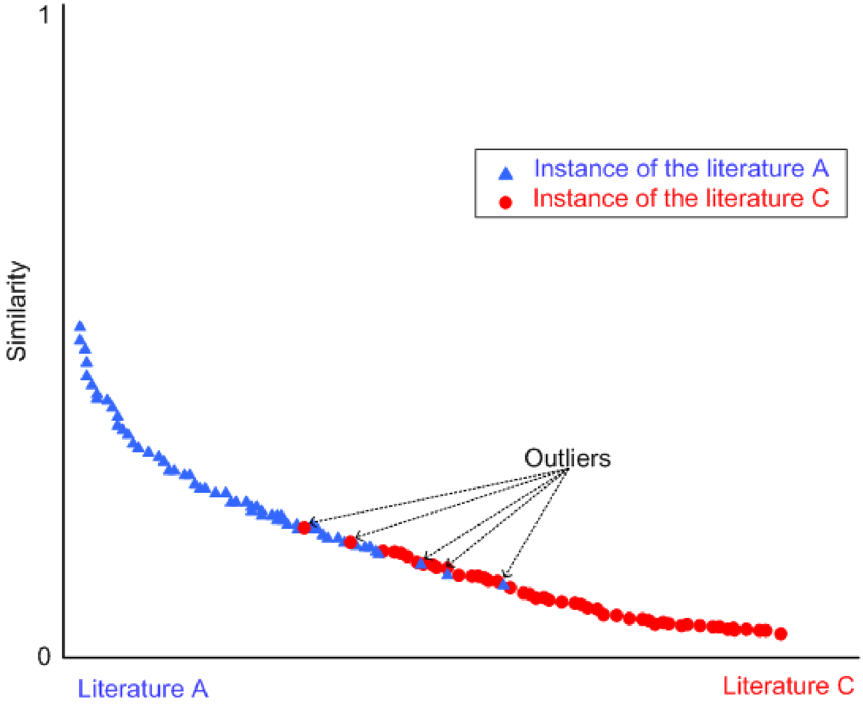


Fig. 1. A graph representing instances (documents) of literature *A* and instances (documents) of literature *C* according to their content similarity to a prototypical document of literature *A*. In this similarity graph, outliers of literature *C* are positioned closer to the typical representatives of the literatures *A* than to the central documents of literature *C*.

Each document from the two literatures is an instance, represented by a set of words using frequency statistics based on the Bag of Words (BoW) text representation [26]. The BoW vector enables to measure content similarity of documents. Content similarity computation is performed with OntoGen, which was designed for interactive data-driven construction of topic ontologies [25]. Content similarity is measured using the standard TF*IDF (term frequency inverse document frequency) weighting method [27], where high frequency of co-occuring words in documents indicates high document similarity. The similarity between documents is visualized with OntoGen in the document's similarity graph, as illustrated in Figure 1.

The cosine similarity measure, commonly used in information retrieval and text mining to determine the semantic closeness of two documents where document features are represented using the BoW vector space model, is used to position the documents according to their similarity to the representative document (centroid) of a selected domain. Documents positioned based on the cosine similarity measure can be visualized in OntoGen by a similarity graph with cosine similarity values that fall within the [0, 1] interval. Value 0 means extreme dissimilarity, where two documents (a given document and the centroid vector of its cluster) share no common words, while value 1 represents the similarity between two semantically identical documents in the BoW representation.

The method uses domains *A* and *C*, and builds a joint document set *AC* (i.e. A∪C). For this intention, two individual sets of documents (e.g. titles, abstracts or full texts of scientific articles), one for each term under research (namely, literature *A* and literature *C*), are automatically retrieved from bibliographic databases or extracted from other document sources. The documents from the two individual sets are loaded as a single text file (i.e. a joint document set *AC*) where each line represents a document with the first word in the line being its name. We consider all the terms and not just the medical ones. A list of 523 English stop words is then used to filter out meaningless words, and English Porter stemming is applied.

From a joint document set A∪C, a similarity graph (Figure 1) between two document sets *A* and *C* is constructed with OntoGen by ranking and visualizing all the documents from *AC* in terms of their similarity to centroid *a* of document set *A*. The OntoGen tool can then be used to build two document clusters, *A'* and *C'* (where A'∪C'=*AC*) in an unsupervise manner, using OntoGen's 2-means clustering algorithm. Cluster *A'* consists mainly of documents from *A,* but may contain also some documents from *C*. Similarly, cluster *C'* consists mainly of documents from *C,* but may contain also some documents from *A*.

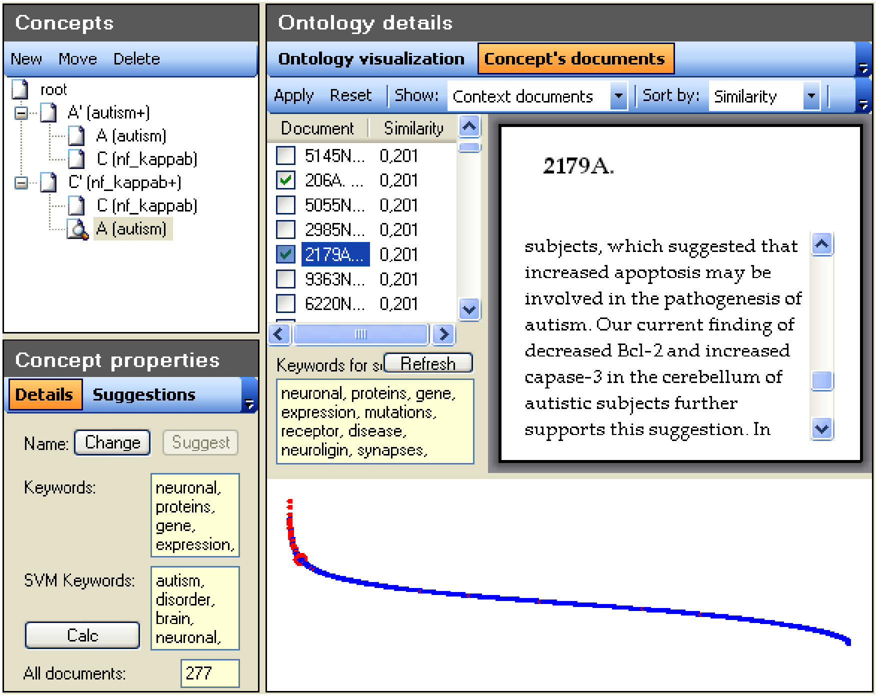
Each cluster is further divided into two document subclusters based on domains *A* and *C* with the aim to identify outlying documents. For each individual document cluster we proceed as follows: cluster *A'* is divided into subclusters A'∩A and A'∩C, while cluster *C'* is divided into C'∩A and C'∩C. In this manner, subclusters A'∩C (outliers of *C*, consisting of documents of domain *C* only) and C'∩A (outliers of *A*, consisting of documents of domain *A* only) are the two document sets that consist of outlying documents.

5 Application of Outlier Detection in the Autism Literature

This section is dedicated to a practical application of the upgraded RaJoLink methodology to text analysis of biomedical scientific documents. We present how text mining and link analysis techniques, which are implemented in our approach, can be performed 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. When exploring the literature on autism, the collaborating medical expert has proposed to take the NF-kappaB literature as one of the most promising potential target domains for further focused studies [7]. For a given hypothesis of NF-kappaB and autism relationship we automatically extracted abstracts 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 group of articles on autism. Technically speaking, when autism literature was selected as domain *A* and when from the joint domain *AC* (joining autism and NF-kappaB literatures) OntoGen's 2-means clustering method was applied to obtain document groups (clusters) *A'* and *C'*, some documents from domain C (NF-kappaB literature) appeared as members of document cluster *A'* containing mostly articled from domain *A* (autism). Similarly, there were also some articles on autism in the group of articles on NF-kappaB (article group *C'*). It turned out that indeed these exceptional documents contain uncommon and therefore potentially bridging terms. In particular, terms Bcl-2, cytokines, MCP-1, oxidative stress and other meaningful linking terms between the literature on autism and the literature on NF-kappaB were detected in these outlier documents.

Here we present finding of an abstract of MEDLINE articles that makes logical connection between the specific autism observations and the NF-kappaB findings across the bridging term Bcl-2, a regulatory protein for control of programmed brain cell death. Figure 2 shows the similarity graph representing instances of literature *A* (autism context) among instances of literature *C* (nf-kappab+ context) according to their content similarity, where *A* denotes a set of documents containing term autism, *A'* denotes the group of documents constructed from the *AC* document set where most documents are documents on autism (i.e., the so-called autism+ context, where + autism being the majority document class in this document group), and *C* denotes a set of documents containing term NF-kappaB (i.e., the so-called nf-kappab+ context).

The presented bisociative linking approach suggests a novel way to improve the evidence gathering phase when analyzing individual terms appearing in literature *A* in terms of their potential for connecting with terms from literature *C*. In fact, even



root
A' (autism+)
A (autism)
C (nf_kappab)
C' (nf_kappab+)
C (nf_kappab)
A (autism)

Show: Context documents
Sort by: Similarity

Document	Similarity
□ 5145N...	0,201
☑ 206A. ...	0,201
□ 5055N...	0,201
□ 2985N...	0,201
☑ 2179A...	0,201
□ 9363N...	0,201
□ 6220N...	0,201

neuronal, proteins, gene, expression, mutations, receptor, disease, neuroligin, synapses,

2179A.

subjects, which suggested that increased apoptosis may be involved in the pathogenesis of autism. Our current finding of decreased Bcl-2 and increased capase-3 in the cerebellum of autistic subjects further supports this suggestion. In

Name:

Keywords: neuronal, proteins, gene, expression,

SVM Keywords: autism, disorder, brain, neuronal,

All documents: 277

Fig. 2. OntoGen's similarity graph representing instances of literature *A* (*autism+ context*) among instances of the literature *C* (*nf-kappab+ context*) according to their content similarity. The distinctive article about the substance Bcl-2 in relation to autism (*2179A*) is visualized among the nf-kappab+ context documents.

Srinivasan and colleagues, who declared to have developed the algorithms that require the least amount of manual work in comparison with other studies [16], still need significant time and human effort for collecting evidence relevant to the hypothesized connections. In the comparable upgraded RaJoLink approach, the domain expert should be involved only in the conclusive actions of the *Link* step to accelerate the choice of significant linking terms. In this step, similarity graph visualization proves to be extremely beneficial for speeding the process of discovering the bridging concepts. Not only that the documents detected as outliers are visualized and their contents presented on the screen by simply clicking on the pixel representing the document (see Figure 2), but also the keywords are listed, explicitly indicating a set of potential bridging concepts (terms) to be explored by the domain experts.

6 Conclusions

Current literature-based approaches depend strictly on simple, associative information search. Commonly, literature-based association is computed using measures of similarity or co-occurrence. Because of their 'hard-wired' underlying criteria of co-occurrence or similarity, these methods often fail to discover relevant information, which is not related in obvious associative ways. Especially information related across separate contexts is hard to identify with the conventional associative approach. In such cases the context-crossing connections, called bisociations, can help generate creative and innovative discoveries. The RaJoLink method has the potential for bisociative relation discovery as it allows switching between contexts and for discovering interesting terms in the intersections between contexts.

Similar to Swanson's closed discovery approach [10], the search for bridging terms consists of looking for terms *b* that can be found in the intersection of two separate sets of records, namely in the literature *A* as well as in the literature *C*. However, our focusing is on outliers from the two sets of records and their neighbouring documents. Thus we show how outlying documents in the similarity graphs yield useful information in the closed discovery, where bridging concepts have to be found between the literatures *A* and *C*. In fact, such visual analysis can show direction to the previously unseen relations like bisociations, which provide new knowledge. This is an important aspect and significant contribution of our method to literature-based discovery research.

Most of the data analysis research is focused on discovering mainstream relations. These relations are well statistically supported; findings usually confirm the conjectured hypothesis. However, this research provides insight into the relationship between outliers and the literature-based knowledge discovery. An important feature of our approach is the way of detecting the bridging concepts connecting unrelated literatures, which we have performed by the OntoGen's similarity graphs. We used them for representing instances of the literature *A* together with instances of the literature *C* according to their content similarity with the goal to identify outliers from the two sets of literatures and their neighbouring documents. 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. Therefore, to test whether the hypothetical observation could be related to the phenomenon under investigation or not, we compare the sets of literature about the initial phenomenon with the literature about the hypothetically related one in the documents' similarity graphs. By our original discovery of linking terms between the literature on autism and the literature on calcineurin we proved that such combination of two previously unconnected sets of literatures in a single content similarity graph can be very effective and useful [5] and [6]. In the autism domain we also discovered a relation between autism and transcription factor NF-kappaB, which has been evaluated by a medical expert as relevant for better understanding of autism [7]. From the similarity graphs that we drew with OntoGen we could quickly notice, which documents from the observed domain are semantically more related to another context. They were positioned in the middle portions of the similarity curves. In the present autism experiment we found a document about the anti-apoptotic protein Bcl-2 [28] that presents a bridging concept among disjoint sets of scientific articles about autism on one hand, and NF-kappaB on the other hand. In fact, Sheikh and colleagues [28] found reduction of Bcl-2, the important marker of apoptosis, in the cerebellum of autistic subjects. Some years before them also Araghi-Niknam and Fatemi showed the reduction of Bcl-2 in superior frontal and cerebellar cortices of autistic individuals [29]. On the other hand, Mattson [30] 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. However, further research about timing, maturational differences in brain development, and other determinants of NF-kappaB involvement in autism would be needed to substantiate the hypotheses generated by our literature-based experiments.

Acknowledgments. This work was partially supported by the project Knowledge Technologies (grant P2-0103) funded by the Slovene National Research Agency, and the EU project FP7-211898 BISON (Bisociation Networks for Creative Information Discovery).

References

1. Moore, D.S., McCabe, G.P.: Introduction to the Practice of Statistics, 3rd edn. W.H. Freeman, New York (1999)
2. Berthold, M.R. (ed.): Bisociative Knowledge Discovery, 1st edn. LNCS(LNAI), vol. 7250. Springer, Heidelberg (2012)
3. Koestler, A.: The act of creation. MacMillan Company, New York (1964)
4. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier detection in cross−context link discovery for creative literature mining. Comput. J., 15 (2010)

5. Petrič, I., Urbančič, T., Cestnik, B., Macedoni−Lukšič, M.: Literature mining method RaJoLink for uncovering relations between biomedical concepts. J. Biomed. Inform. 42(2), 219–227 (2009)
6. Petrič, I., Urbančič, T., Cestnik, B.: Discovering hidden knowledge from biomedical literature. Informatica 31(1), 15–20 (2007)
7. 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. LNCS (LNAI), vol. 4594, pp. 217–226. Springer, Heidelberg (2007)
8. Mednick, S.A.: The associative basis of the creative process. Psychol. Rev. 69(3), 220–232 (1962)
9. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 4th edn. Text Revision, Washington, DC (2000)
10. Swanson, D.R.: Undiscovered public knowledge. Library Quarterly 56(2), 103–118 (1986)
11. 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. J. Am. Soc. Inf. Sci. Tech. 52(7), 548–557 (2001)
12. Smalheiser, N.R., Swanson, D.R.: Using ARROWSMITH: a computer−assisted approach to formulating and assessing scientific hypotheses. Comput. Methods Programs Biomed. 57(3), 149–153 (1998)
13. Swanson, D.R., Smalheiser, N.R., Torvik, V.I.: Ranking indirect connections in literature−based discovery: The role of Medical Subject Headings (MeSH). J. Am. Soc. Inf. Sci. Tech. 57(11), 1427–1439 (2006)
14. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature−based discovery to identify disease candidate genes. Int. J. Med. Inform. 74(2-4), 289–298 (2005)
15. Weeber, M.: Drug Discovery as an Example of Literature-Based Discovery. In: Džeroski, S., Todorovski, L. (eds.) Computational Discovery 2007. LNCS (LNAI), vol. 4660, pp. 290–306. Springer, Heidelberg (2007)
16. Srinivasan, P., Libbus, B.: Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics 20(suppl. 1), I290–I296 (2004)
17. Yetisgen−Yildiz, M., Pratt, W.: Using statistical and knowledge−based approaches for literature−based discovery. J. Biomed. Inform. 39(6), 600–611 (2006)
18. 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)
19. Principles of MEDLINE Subject Indexing, `http://www.nlm.nih.gov/bsd/disted/mesh/indexprinc.html`
20. Lindsay, R.K., Gordon, M.D.: Literature−based discovery by lexical statistics. J. Am. Soc. Inf. Sci. 50(7), 574–587 (1999)
21. Ohsawa, Y.: Chance discovery: the current states of art. Chance Discoveries in Real World Decision Making 30, 3–20 (2006)
22. Magnani, L.: Chance Discovery and the Disembodiment of Mind. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 547–553. Springer, Heidelberg (2005)
23. Grobelnik, M., Mladenić, D.: Extracting human expertise from existing ontologies. EU−IST Project IST−2003−506826 SEKT (2004)
24. MEDLINE Fact Sheet, `http://www.nlm.nih.gov/pubs/factsheets/medline.html`

25. 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.) Proceedings of the 9th International Multi-Conference Information Society, Ljubljana, Slovenia, pp. 223–226 (2006)
26. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (2002)
27. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
28. Sheikh, A.M., Li, X., Wen, G., Tauqeer, Z., Brown, W.T., Malik, M.: Cathepsin D and apoptosis related proteins are elevated in the brain of autistic subjects. Neuroscience 165(2), 363–370 (2010)
29. 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), 945–952 (2003)
30. Mattson, M.P.: NF−kappaB in the survival and plasticity of neurons. Neurochemical Research 30(6-7), 883–893 (2005)