Exploring the Power of Outliers for Cross-Domain Literature Mining

Borut Sluban[1], Matjaž Juršič[1], Bojan Cestnik[1,2], and Nada Lavrač[1,3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Temida d.o.o., Ljubljana, Slovenia
[3] University of Nova Gorica, Nova Gorica, Slovenia
{borut.sluban,matjaz.jursic,bojan.cestnik,nada.lavrac}@ijs.si

Abstract. In bisociative cross-domain literature mining the goal is to identify interesting terms or concepts which relate different domains. This chapter reveals that a majority of these domain bridging concepts can be found in outlier documents which are not in the mainstream domain literature. We have detected outlier documents by combining three classification-based outlier detection methods and explored the power of these outlier documents in terms of their potential for supporting the bridging concept discovery process. The experimental evaluation was performed on the classical migraine-magnesium and the recently explored autism-calcineurin domain pairs.

1 Introduction

Scientific literature serves as the basis of research and discoveries in all scientific domains. In literature-based creative knowledge discovery one of the interesting goals is to identify terms or concepts which relate different domains, as these terms may represent germs of new scientific discoveries.

The aim of this chapter[1] is to present an approach which supports scientists in their creative knowledge discovery process when analyzing scientific papers of their interest. The presented research follows Mednick's *associative creativity theory* [9] defining creative thinking as the capacity of generating new combinations of distinct associative elements (e.g. words), and Koestler's book *The act of creation* [7] stating that scientific discovery requires creative thinking to connect seemingly unrelated information. Along these lines, Koestler explores domain-crossing associations, called *bisociations*, as a crucial mechanism for progressive insights and paradigm shifts in the history of science.

Based on the definition of bisociations—defined by Koestler [7] and further refined by Dubitzky et al. [3]—our work addresses the task of supporting the search for bisociative links that cross different domains. We consider a simplified setting, where a scientist has identified two domains of interest (two different scientific areas or two different contexts) and tries to find concepts that represent potential links between the two different contexts. This simplified cross-context

[1] This chapter is an extension of our short paper [16].

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 325–337, 2012.

link discovery setting is usually referred to as the *closed discovery* setting [23]. Like Swanson [19] and Weeber et al. [23], we address the problem of literature mining, where papers from two different scientific areas are available, and the task is to support the scientist in cross-context literature mining. By addressing this task, our aim is to contribute to a methodology for semi-automated cross-context literature mining, which will advance both the area of computational creativity as well as the area of text mining.

We investigate the role of *outliers* in literature mining, and explore the utility of outliers in this non-standard text mining task of *cross-context link discovery*. We provide evidence that outlier detection methods can contribute to literature-based cross-domain scientific discovery based on the notion of bisociation.

This chapter is organized as follows. Section 2 presents the related work in literature mining and outlier detection. In Section 3 we present the experimental datasets, and the method for transforming a set of documents into a format required for text processing and outlier detection. Section 4 presents the methodology for outlier document detection in cross-domain knowledge discovery, together with its evaluation in two medical problem settings: in the classical migraine-magnesium cross-domain discovery problem and in the autism-calcineurin domain pair. Section 5 concludes by a discussion and directions for further work.

2 Related Work

The motivation for new scientific discoveries from disparate literature sources grounds in Mednick's *associative creativity theory* [9] and in the literature on domain-crossing associations, called *bisociations*, introduced by Koestler [7]. Furthermore, we are inspired by the work of Weeber et al. [23] who followed the work of creative literature-based discovery in medical domains introduced by Swanson [19]. Swanson designed the *ABC model* approach that investigates whether an agent A is connected with a phenomenon C by discovering complementary structures via interconnecting phenomena B (see Figure 1)[2]. 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 may contribute to a better understanding of phenomenon C.

In a *closed discovery process*, where domains A and C are specified by the expert at the beginning of the discovery process, the goal is to search for bridging concepts (terms) b in B in order to support the validation of the hypothesized connection between A and C (see Figures 1 and 2). Smalheiser and Swanson [17] developed an online system ARROWSMITH, which takes as input two sets of titles from disjoint domains A and C and lists b-terms that are common to literature A and C; the resulting bridging terms (b-terms) are used to generate

[2] Uppercase letter symbols A, B and C are used to represent sets of terms, and lowercase symbols a, b and c to represent single terms.

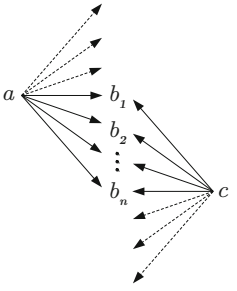

Fig. 1. Closed discovery process as defined by Weeber et al. [23]

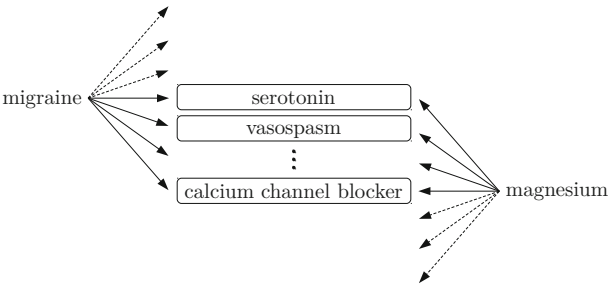

Fig. 2. Closed discovery when exploring migraine and magnesium documents, with *b*-terms as identified by Swanson et al. [21]

novel scientific hypotheses. As stated by Swanson et al. [21], the major focus in literature-based discovery has been on the closed discovery process, where both A and C are specified in advance.

Srinivasan [18] developed an algorithm for bridging concept identification that is claimed to require the least amount of manual work in comparison with other literature-based discovery studies. However, it still needs substantial time and human effort for collecting evidence relevant to the hypothesized connections. In comparison, one of the advantages of the approach presented in this chapter is that the domain expert needs to be involved only in exploring the potential *b*-terms in outlier documents, instead of exploring all the most frequent potential *b*-terms in all the documents.

In a closely related approach, rarity of terms as means for knowledge discovery has been explored in the RaJoLink system [13,22], which can be used to find interesting scientific articles in the PubMed database with the aim to discover new knowledge. The 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, respectively. In the Ra step, interesting rare terms in literature about the phenomenon A under investigation are identified. In the Jo step, 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 as the candidates for C. This results in a candidate hypothesis that C is connected with A. In order to provide explanation for hypotheses generated in the Jo step, in the Link step the method searches for *b*-terms, linking the literature on joint term c from C and the literature on term a from A. Note that steps Ra and Jo implement the open discovery, while step Link corresponds to the closed discovery process, searching for *b*-terms when A and C are already known, as illustrated in Figure 1. Figure 2 illustrates the closed discovery process for a real-life case of exploring the migraine-magnesium domain pair.

Petrič et al. [10] have recently upgraded the RaJoLink methodology by inspecting outlier documents as a source for speeding up the *b*-term detection process. Like in this work—and similar to the definition of outliers in statistics where an outlier is defined as an observation that is numerically distant from the rest of the data—they also focus on outlier observations (documents) that lie outside the overall pattern of the given (class) distribution. More specifically, their methodology focuses on the search for *b*-terms in outlier documents identified by OntoGen, a semi-automated tool for topic ontology construction [4]. Opposed to their approach, which uses *k*-means clustering in OntoGen to detect outlier documents included in the opposite cluster [14], our approach uses several classification algorithms to identify misclassified documents as domain outliers, which are inspected for containing domain bridging terms.

Since outlier mining has already proved to have important applications in fraud detection and network intrusion detection [1], we focused on outliers as they may actually have the potential to lead to the discovery of intriguing new information. Classification noise filters and their ensembles, recently investigated by the authors [15], are used for outlier document detection in this chapter. Documents of a *domain pair* dataset (i.e., the union of two different domain literatures) that are misclassified by a classifier can be considered as domain outliers, since these instances tend to be more similar to regular instances of the opposite domain than to instances of their own domain. The utility of domain outliers as relevant sources of domain bridging terms is the topic of study of this chapter.

3 Experimental Datasets

This section shortly describes two datasets which were used to evaluate the proposed outlier detection approach for cross-domain literature mining. Along with the descriptions of datasets we also provide the description of our preprocessing techniques and some basic statistics for the reader to get a better idea of the data.

The first dataset - the *migraine-magnesium* domain pair - was previously well researched by different authors [13,19,20,21,23]. In the literature-based discovery process Swanson managed to find more than 60 pairs of articles connecting the migraine domain with the magnesium deficiency via several bridging concepts. In this process Swanson identified 43 *b*-terms connecting the two domains of the *migraine-magnesium* domain pair [21].

The second dataset - the *autism-calcineurin* domain pair - was introduced and initially researched by Petrič et al. in [11,12,22] and later also in [8,10,13]. 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. It is a very complex and not yet sufficiently understood domain, where precise causes are still unknown. Alike Swanson, Petrič et al. [13] also provide *b*-terms, 13 in total, whose importance in connecting autism to calcineurin (a protein phosphatase) is discussed and confirmed by the domain expert.

Table 1. Bridging terms – *b*-terms identified by Swanson et al. [21] and Petrič et al. [13] for the *migraine-magnesium* and *autism-calcineurin* domain pair, respectively

migraine-magnesium	**autism-calcineurin**
serotonin, spread, spread depression, seizure, calcium antagonist, vasospasm, paroxysmal, stress, prostaglandin, reactivity, spasm, inflammatory, anti inflammatory, 5 hydroxytryptamine, calcium channel, epileptic, platelet aggregation, verapamil, calcium channel blocker, nifedipine, indomethacin, prostaglandin e1, anticonvulsant, arterial spasm, coronary spasm, cerebral vasospasm, convulsion, cortical spread depression, brain serotonin, 5 hydroxytryptamine receptor, epilepsy, antimigraine, 5 ht, epileptiform, platelet function, prostacyclin, hypoxia, diltiazem, convulsive, substance p, calcium blocker, prostaglandin synthesis, anti aggregation	synaptic, synaptic plasticity, calmodulin, radiation, working memory, bcl 2, type 1 diabetes, ulcerative colitis, asbestos, deletion syndrome, 22q11 2, maternal hypothyroxinemia, bombesin

We use the *b*-terms, which were identified in each of the two domain pair datasets, as the gold standard to evaluate the utility of domain outlier documents in the cross-context link discovery process. Table 1 presents the *b*-terms for the *migraine-magnesium* and the *autism-calcineurin* domain pair datasets used in our experiments.

Both datasets were retrieved from the PubMed database[3] using the keyword query; however, we used additional filtering condition for selection of migraine-magnesium dataset. It was necessary to select only the articles published before the year 1988 as this was the year when Swanson published his research about this dataset and thus making an explicit connection between migraine and magnesium domain. Preprocessing was done in a standard text mining way, using the preprocessing steps described in [6]: (a) text tokenization, (b) stopword removal, (c) word stemming/lemmatization using LemmaGen lemmatizer for English [5], (d) construction of N-grams which are terms defined as a concatenation of 1 to N words than appear consecutively in text with minimum supporting frequency, (e) creation of standard bag-of-words (BoW) representation of text using term-frequency-inverse-document-frequency (tf-idf) or binary (depends on classification algorithm) term weights. Besides this standard workflow we additionally removed from the dataset all terms (N-grams) containing words which were used as query terms during document selection. Experiments showed that the correlation between the domain class and the query terms is too high for an outlier detection algorithm to find a reasonable number of high quality outliers. A summary of statistics on the datasets used in our experiments is presented in Table 2.

The 43 *b*-terms identified by Swanson in the standard *migraine-magnesium* dataset were retrieved from article titles only [21]. Therefore, we also used only article titles in our experiments. In the preprocessing of this dataset we constructed 3-grams to obtain more features for each document despite a relatively

[3] PubMed: `http://www.ncbi.nlm.nih.gov/pubmed`

Table 2. Some basic properties and statistics of the domain pair datasets used in the experimental evaluation

Dataset name	migraine-magnesium	autism-calcineurin
Document source	PubMed	PubMed
Query terms	"migraine" |"magnesium" (condition: year<1988)	"autism" |"calcineurin"
Number of retrieved doc.	8,058 (2,425 |5,633)	15,243 (9,365 |5,878)
Part of document used (text)	title	abstract
Average text length	11 words, 12 terms	180 words, 105 terms
Term definition	3-grams, min. freq. 2	1-grams, min. freq. 15
Number of distinct terms	13,524	5,255
Number of *b*-terms	43	13
Num. of doc. with *b*-terms	394 = 4.89%	1672 = 10.97%

low average word count. On the other hand, for the *autism-calcineurin* dataset, which contains titles and the abstracts, we had to limit ourselves to 1-grams and had to set the minimum supporting frequency of terms higher to reduce the number of features due to computational limitations.

4 Detecting Outlier Documents

This research aims at supporting the search for cross-domain links between concepts from two disparate literatures A and C, based on exploring outlier articles of the two domains. Our method assumes that by exploring outlier documents it will be easier to discover linking b-terms (bridging concepts) that establish previously unknown links between literature A and literature C, as the hypothesis of this work is that most bridging concepts occur in outlier documents. This section first presents the algorithms used for outlier detection, followed by the experimental validation of our hypothesis that outlier documents contain a relatively higher number of bridging terms than other documents.

4.1 Classification Noise Filters for Outlier Detection

The novelty of our work is to use noise detection approaches for findinging outlier documents containing cross-domain links (bridging terms – b-terms) between different domains. When exploring a domain pair dataset we searched for a set of outlier documents with different classification noise filtering approaches [2], implemented and adapted for this purpose.

Classification noise filtering is based on the idea of using a classifier as a tool for detecting noisy and outlier instances in data. In this work the simple classifiers used in [2] were replaced by new, better performing classifiers, as the noise filter should, as much as possible, trust the classifiers that they will be able to correctly predict the class of a data instance. In this way the incorrectly classified instances are considered to be noise/outliers. In other words, if an instance of class A is classified in the opposite class C, we consider it to be an

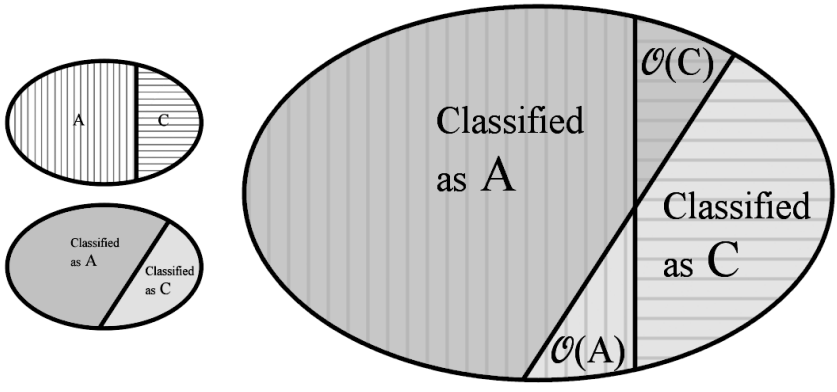

Fig. 3. Detecting outliers of a domain pair dataset with classification filtering

outlier of domain A, and vice versa. We denote the two sets of domain outlier documents with *O(A)* and *O(C)*, respectively. Figure 3 depicts this principle.

The proposed outlier detection method works in a 10-fold cross-validation manner, where repeatedly nine folds are used for training the classifier and on the complementary fold the misclassified instances are denoted as noise/outliers. Instances of a domain pair dataset that are misclassified by a classifier can be considered as domain outliers, since these instances tend to be more similar to regular instances of the opposite domain than to instances of their own domain.

4.2 Experimental Evaluation

The goal of this section is to provide experimental evidence for the hypothesis that outliers can be used as the focus of exploration to speed-up the search for bridging concepts between different domains of expertise [10,14]. Therefore, our experiments are designed to validate that sets of outlier documents are rich on *b*-terms and contain significantly more *b*-terms than sets of arbitrary documents.

We implemented three classification noise detection algorithms, using three different classifiers: Naïve Bayes (abbreviated: Bayes), Random Forest (RF) and Support Vector Machine (SVM). In addition to the outlier sets obtained by these three classification filters, we examined also the union of these outlier sets and the so called "Majority" outlier set containing outlier documents that were detected by at least two out of three classification filters.

Our experiments were performed on the migraine-magnesium and the autism-calcineurin domain pair datasets, described in Section 3. To measure the relevance of the detected outlier documents in terms of their potential for containing domain bridging terms, we inspected 43 terms known as bridging terms appearing in the migraine-magnesium domain pair and 13 known *b*-terms in the autism-calcineurin domain pair. Tables 3 and 4 present the size of all examined sets of outlier documents and the amount of *b*-terms they contain, for the migraine-magnesium and autism-calcineurin dataset, respectively.

Table 3. Numbers of documents and *b*-terms (bT) in different outlier sets, with percentages showing their proportion compared to all documents in the migraine-magnesium domain. The bT percentages can be interpreted as recall of *b*-terms.

Class	All docs		Bayes		RF		SVM		Union		Majority	
	Docs	bT	Docs	bT	Docs	bT	Docs	bT	Docs	bT	Docs	bT
MIG	2,425	43	248	23	772	26	192	12	895	34	237	20
	(100%)	(100%)	(10%)	(53%)	(32%)	(60%)	(8%)	(28%)	(37%)	(79%)	(10%)	(47%)
MAG	5,633	43	335	27	124	19	170	24	475	33	131	21
	(100%)	(100%)	(6%)	(63%)	(2%)	(44%)	(3%)	(56%)	(8%)	(77%)	(2%)	(49%)
Total	8,058	43	**583**	**32**	**896**	**36**	**362**	**29**	**1,370**	**40**	**368**	**30**
	(100%)	(100%)	**(7%)**	**(74%)**	**(11%)**	**(84%)**	**(4%)**	**(67%)**	**(17%)**	**(93%)**	**(5%)**	**(70%)**
Randomly sampled	8,058	43	**583**	**19**	**896**	**24**	**362**	**14**	**1,370**	**29**	**368**	**14**
	(100%)	(100%)	**(7%)**	**(44%)**	**(11%)**	**(56%)**	**(4%)**	**(33%)**	**(17%)**	**(68%)**	**(5%)**	**(33%)**

Table 4. Numbers of documents and *b*-terms (bT) in different outlier sets, with percentages showing their proportion compared to all documents in the autism-calcineurin domain. The bT percentages can be interpreted as recall of *b*-terms.

Class	All docs		Bayes		RF		SVM		Union		Majority	
	Docs	bT	Docs	bT	Docs	bT	Docs	bT	Docs	bT	Docs	bT
AUT	9,365	13	349	8	77	7	147	6	388	9	139	7
	(100%)	(100%)	(4%)	(62%)	(1%)	(54%)	(2%)	(46%)	4%)	(69%)	(1%)	(54%)
CAL	5,878	13	316	7	65	5	145	7	397	10	98	6
	(100%)	(100%)	(5%)	(54%)	(1%)	(38%)	(2%)	(54%)	(7%)	(77%)	(2%)	(46%)
Total	15,243	13	**665**	**9**	**142**	**9**	**292**	**9**	**785**	**12**	**237**	**9**
	(100%)	(100%)	**(4%)**	**(69%)**	**(1%)**	**(69%)**	**(2%)**	**(69%)**	**(5%)**	**(92%)**	**(2%)**	**(69%)**
Randomly sampled	15,243	13	**665**	**8**	**142**	**5**	**292**	**6**	**785**	**9**	**237**	**6**
	(100%)	(100%)	**(4%)**	**(63%)**	**(1%)**	**(35%)**	**(2%)**	**(46%)**	**(5%)**	**(67%)**	**(2%)**	**(46%)**

Columns of Tables 3 and 4 present the numbers of outlier documents (and contained *b*-terms) identified by different outlier detection approaches, together with percentages showing their proportion compared to the given dataset. The rows present these numbers separately for each class, for both classes together, and—for the needs of results validation explained below—for a random sample of documents in the size of the detected outlier set.

These results show that all five outlier subsets[4] of each of the two domain pairs contain from 70% to over 90% (for the "Union" subset) of *b*-terms, on average in less than 10% of all documents from the migraine-magnesium dataset and in less than 5% of all documents of the autism-calcineurin dataset. This means that by inspecting outlier documents, which represent only a small fraction of the datasets, a great majority of *b*-terms can be found, which substantially reduces the time and effort needed by the domain expert to discover cross-domain links.

To confirm that these results are not due to chance (do not hold for just any arbitrary subset that has the same size as an outlier set), we have randomly sampled 1,000 subsets for each of the five outlier sets (all of them having the same size as their corresponding outlier set) in order to present the average *b*-term occurrences in randomly sampled subsets. The last row of Tables 3 and 4 shows that the sets of outlier documents contain on average more than 30% more of all *b*-terms in the migraine-magnesium dataset and more than 20% more of all *b*-terms in the autism-calcineurin dataset than randomly sampled sets of the same size.

A comparison of the above discussed results relative to the whole migraine-magnesium and autism-calcineurin datasets is summarized in Figures 4 and 5, respectively.

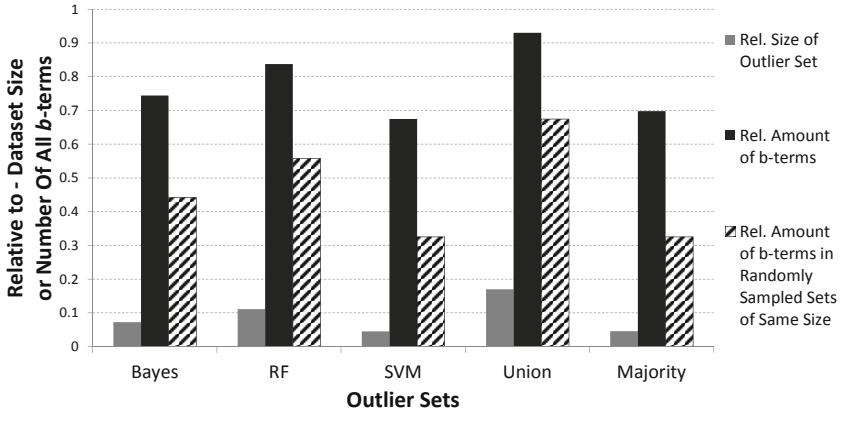

Fig. 4. Relative size of outlier sets and the amount of *b*-terms for the migraine-magnesium dataset

[4] *Outlier subset* is used instead of *outlier set* to emphasize its relation to the entire dataset of documents. The terms are used interchangeably, however they always refer to a set of detected outlier documents that belong to a certain domain pair.

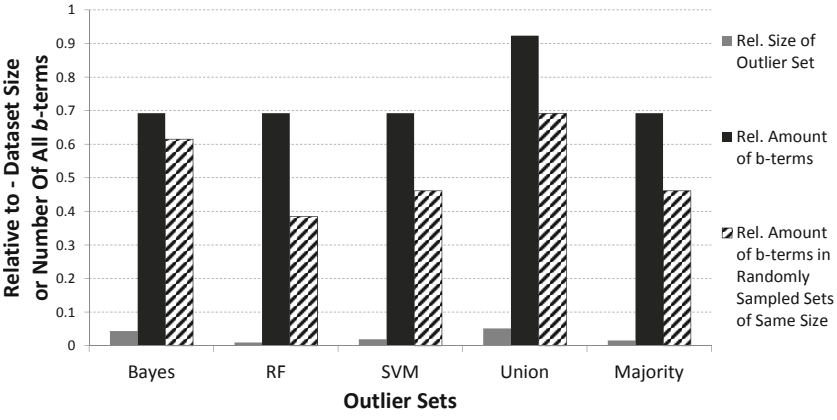

Fig. 5. Relative size of outlier sets and the amount of *b*-terms for the autism-calcineurin dataset

Additionally, we compared relative frequencies of *b*-terms in the detected outlier sets to their relative frequencies in the whole dataset, i.e. the fraction of documents containing a certain *b*-term among the documents of a chosen set. In Figure 6 we present the increase of relative frequencies of *b*-terms in the "Majority" outlier set detected on the migraine-magnesium dataset.

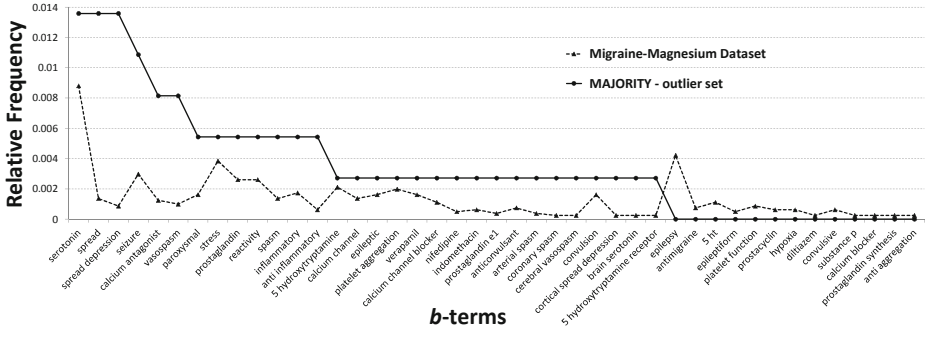

Fig. 6. Comparison of relative frequencies of bridging terms in the entire migraine-magnesium dataset and in the "Majority" set of outlier documents detected by three different outlier detection methods

The "Majority" outlier set approach proved to have the greatest potential for bridging concept detection. Firstly, because of the best ratio among the proportion of the size of the outlier subset and the proportion of *b*-terms which are present in that outlier subset (see Table 4 and Figure 4), and secondly, because the relative frequency of all the *b*-terms present in the"Majority" outlier set is higher compared to the entire migraine-magnesium dataset, as can be clearly seen from Figure 6.

Similarly, encouraging results for the "Majority" outlier set detected on the autism-calcineurin dataset can be observed in Figure 7.

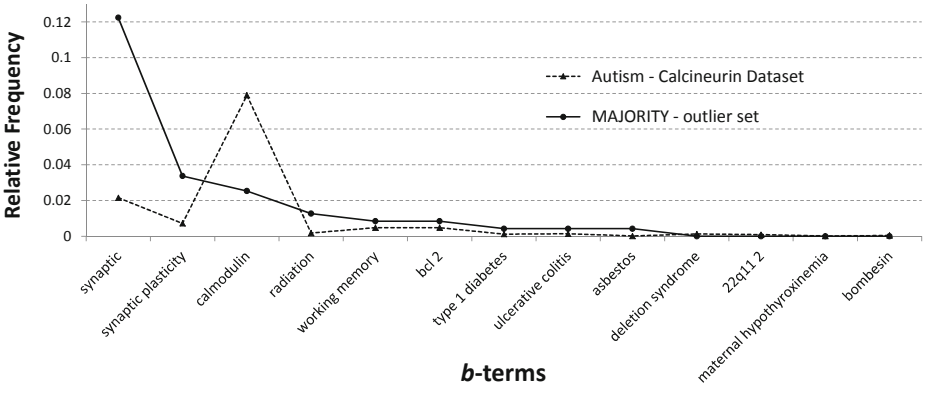

Fig. 7. Comparison of relative frequencies of bridging terms in the entire autism-calcineurin dataset and in the "Majority" set of outlier documents detected by three different outlier detection methods[5]

All *b*-terms that are present in the "Majority" outlier set, except for one ("*calmodulin*"), have a higher relative frequency in the outlier set compared to the relative frequency in the entire dataset. Although (1) the RF outlier set is best in terms of the ratio among the proportion of the size of the outlier subset and the proportion of *b*-terms which are present in that outlier subset and (2) the "Majority" outlier set is second best (for the autism-calcineurin dataset), in general we prefer the "Majority" outlier set for bridging concept detection. The majority approach is more likely to give quality outliers on various datasets, in contrast to a single outlier detection approach, since it reduces the danger of overfitting or bias to a certain domain by requiring the agreement of at least two outlier detection approaches for a document to declare it as an domain outlier.

5 Conclusions

In our research we investigated the potential of outlier detection methods in literature mining for supporting the discovery of bridging concepts between disparate domains.

We retrieved articles for the migraine-magnesium and the autism-calcineurin domain pairs from the PubMed database. In our experiments we obtained five sets of outlier documents for each domain pair by three different outlier detection methods, their union and a majority voting approach. Experimental results

[5] Note that the scale of the chart in Figure 7 is different from the scale of the chart in Figure 6.

show that inspecting outlier documents considerably contributes to the bridging concept discovery process, since it enables the expert to focus only on a small fraction of documents which is rich on concept bridging terms (b-terms). Thus, the effort needed for finding cross-domain links is substantially reduced, as it requires to explore a much smaller subset of documents, where a great majority of b-terms are present and more frequent.

In further work we will examine other outlier detection methods in the context of cross-domain link discovery and use outlier documents as a heuristic guidance in the search for potential b-terms on yet unexplored domain-pairs.

Acknowledgement. This work was supported by the European Commission in the context of the 7th Framework Programme FP7-ICT-2007-C FET-Open, contract no. BISON-211898, and in the context of the FP7 project FIRST under the grant agreement no. 257928.

References

1. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: Sellis, T. (ed.) Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, pp. 37–46 (2001)
2. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. Journal of Artificial Intelligence Research 11, 131–167 (1999)
3. Dubitzky, W., Kötter, T., Schmidt, O., Berthold, M.R.: Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 11–32. Springer, Heidelberg (2012)
4. Fortuna, B., Grobelnik, M., Mladenic, D.: OntoGen: Semi-automatic Ontology Editor. In: Smith, M.J., Salvendy, G. (eds.) HCII 2007. LNCS, vol. 4558, pp. 309–318. Springer, Heidelberg (2007)
5. Juršič, M., Mozetič, I., Erjavec, T., Lavrač, N.: Lemmagen: Multilingual lemmatisation with induced ripple-down rules. Journal of Universal Computer Science 16(9), 1190–1214 (2010)
6. Juršič, M., Sluban, B., Cestnik, B., Grčar, M., Lavrač, N.: Bridging Concept Identification for Constructing Information Networks from Text Documents. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 66–90. Springer, Heidelberg (2012)
7. Koestler, A.: The act of creation. MacMillan Company, New York (1964)
8. Macedoni-Lukšič, M., Petrič, I., Cestnik, B., Urbančič, T.: Developing a deeper understanding of autism: Connecting knowledge through literature mining. Autism Research and Treatment (2011)
9. Mednick, S.A.: The associative basis of the creative process. Psychological Review 69, 219–227 (1962)

10. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Outlier detection in cross-context link discovery for creative literature mining. The Computer Journal (2010)
11. Petrič, I., Urbančič, T., Cestnik, B.: Literature mining: Potential for gaining hidden knowledge from biomedical articles. In: Bohanec, M., et al. (eds.) Proceedings of the 9th International Multiconference Information Society, pp. 52–55 (2006)
12. Petrič, I., Urbančič, T., Cestnik, B.: Discovering hidden knowledge from biomedical literature. Informatica 31, 15–20 (2007)
13. 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), 220–232 (2009)
14. Petrič, I., Cestnik, B., Lavrač, N., Urbančič, T.: Bisociative Knowledge Discovery by Literature Outlier Detection. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 313–324. Springer, Heidelberg (2012)
15. Sluban, B., Gamberger, D., Lavrač, N.: Performance analysis of class noise detection algorithms. In: Ågotnes, T. (ed.) Proceedings of the 5th Starting AI Researchers Symposium - STAIRS at ECAI 2010, pp. 303–314 (2011)
16. Sluban, B., Juršič, M., Cestnik, B., Lavrač, N.: Evaluating Outliers for Cross-Context Link Discovery. In: Peleg, M., Lavrač, N., Combi, C. (eds.) AIME 2011. LNCS, vol. 6747, pp. 343–347. Springer, Heidelberg (2011)
17. 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)
18. Srinivasan, P.: Text mining: Generating hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology 55, 396–413 (2004)
19. Swanson, D.R.: Undiscovered public knowledge. Library Quarterly 56(2), 103–118 (1986)
20. Swanson, D.R.: Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association 78(1), 29–37 (1990)
21. 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), 1427–1439 (2006)
22. 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)
23. 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, 548–557 (2001)