Bisociative Literature-Based Discovery: Lessons Learned and New Prospects

Nada Lavrač[1,2], Matej Martinc[1,4], Senja Pollak[1], Bojan Cestnik[3,1]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] University of Nova Gorica, Nova Gorica, Slovenia
[3] Temida d.o.o, Ljubljana, Slovenia
[4] International Postgraduate School Jožef Stefan, Ljubljana, Slovenia

Abstract

The field of bisociative literature-based discovery aims at exploring scientific literature to reveal new discoveries based on yet uncovered relations between knowledge from different, relatively isolated fields of specialization. This paper outlines selected outlier-based literature mining approaches, which focus on finding outlier documents as means for finding unexpected links crossing different contexts. Selected approaches to bridging term detection through outlier document exploration are briefly outlined, together with the lessons learned from recent applications in medical and biological literature-based knowledge discovery. Finally, the paper addresses new prospects in bisociative literature-based discovery, emphasizing the use of advanced embeddings technology for cross-domain literature mining.

Introduction

Understanding complex phenomena and solving difficult problems often requires knowledge from different domains to be combined and cross-domain associations to be considered. While the concept of association is at the heart of several information technologies, including information retrieval and data mining e.g., association rule learning (Agrawal et al., 1996), scientific discovery usually requires to connect seemingly unrelated information from different domains. These kinds of bisociative context crossing associations, called *bisociations* (Koestler, 1964), are often needed for innovative discoveries.

In literature-based discovery (LBD) (Bruza and Weeber, 2008)—and in particular in cross-domain literature mining, which addresses knowledge discovery in two (several) initially separate document corpora—a crucial step is the identification of interesting bridging terms (b-terms) or links (b-links) that carry the potential of explicitly revealing the connections between the separate domains. Swanson (1990) and Smalheiser (1998) developed early LBD approaches to detecting interesting b-terms to uncover the possible cross-domain relations among previously unrelated concepts. For example, the ARROWSMITH online system, developed by Smalheiser and Swanson (1998), takes as input two sets of titles of scientific papers from disjoint domains (disjoint document corpora) A and C, and lists terms that are common to A and C; the resulting bridging terms (b-terms) are further investigated by the user for their potential to generate new scientific hypotheses.[1] Their approach, known as the 'ABC model of knowledge discovery', is based on the *closed discovery* setting (Weeber et al., 2001), where two initially separate domains A and C are specified by the user at the beginning of the discovery process, and the goal is to search for bridging concepts (terms) b in B to validate the hypothesized connection between A and C. The closed discovery setting addressed in this paper is illustrated in Figure 1.

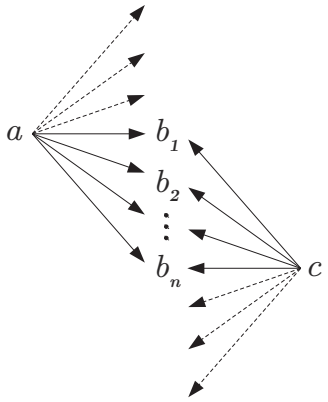

Figure 1: Closed discovery process defined by Weeber et al. (2001).

Inspired by early Swanson's and Smalheiser's work, literature mining approaches were further developed and successfully applied to different problems, such as finding associations between genes and diseases (Hristovski et al., 2005), diseases and chemicals (Yetisgen-Yildiz and Pratt, 2006), and many others. Holzinger et al. (2013) describe several quality-oriented web-based tools for the analysis of biomedical literature, which include the analysis of terms (biomedical entities such as disease, drugs, genes, proteins and organs) and provide concepts associated with a given term. The work of Kastrin, Rindflesch, and Hris-

[1] In the ABC model, uppercase letter symbols A, B and C are used to represent concepts (or sets of terms), and lowercase symbols a, b and c to represent single terms.

Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)
ISBN: 978-989-54160-2-8

tovski (2014) is complementary to other LBD approaches, as it uses different similarity measures (such as common neighbors, Jaccard index, and preferential attachment) for link prediction of implicit relationships in the Semantic MEDLINE network. A comprehensive survey of modern literature-based discovery approaches in biomedical domain can be found in Sebastian, Siew, and Orimaye (2017) and Gopalakrishnan et al. (2019).

This work follows the line of research in two closely related areas, which provide computational tools that act as creative assistants to support human creativity: *literature-based discovery*, described in some detail above, as well as *bisociative knowledge discovery*, where—according to Berthold (2012)—two concepts are bisociated if there is no direct, obvious evidence linking them and if one has to cross different domains to find the link, where a new link must provide some novel insight into the problem addressed. In the context of this paper, both research areas address the same computational creativity task of bridging term (b-term) detection when exploring the connections between two different domains of interest.

More generally, bisociative knowledge discovery addresses a data mining task where two or more domains of interest are searched for bisociative links or individual bridging concepts (i.e. individual context bridging terms). Bisociative knowledge discovery differs from more standard discovery science and associative data mining approaches, like the standard association rule learning (Agrawal et al., 1996), which focus on knowledge discovery within a given domain of interest. Notably, the ability of literature-based discovery and bisociative knowledge discovery methods and software tools that aim to support the experts in their knowledge discovery processes—especially in searching for yet unexplored connections between different domains—is becoming increasingly important.

This paper outlines selected approaches to cross-domain literature mining that support the expert in searching for hidden links connecting two seemingly unrelated domains. The next section below outlines our early approaches to cross-domain literature mining via outlier document detection and exploration (Petrič et al., 2012; Sluban et al., 2012), together with the lessons learned from their past applications in medical literature mining. The subsequent section presents a more recent implementation of the outlier-based approach to LBD (Cestnik et al., 2017), which implements ensemble-based term ranking using an ensemble of six elementary heuristics for b-term evaluation, and incorporates also the human-computer interface (HCI) of the CrossBee b-term detection system (Juršič et al., 2012), together with the lessons learned from the recent LBD applications. The literature based discovery workflow implemented in TextFlows (Perovšek et al., 2016), acting as the enabling technology for implementing the developed cross-domain link discovery approach, is also briefly mentioned. The last sections propose some future research directions based on the lessons learned from the current text mining research, including the idea of a novel LBD framework motivated by the recent word embedding technology. The paper concludes with a summary and some plans for further work.

Outlier-based LBD: Early results and lessons learned

Outliers, characterized by their properties of being infrequent or unusual, may represent unexpected events, entities, items or documents. Early research in LBD has focused on the identification and exploration of outlier documents since they frequently embody new information that is often hard to explain in the context of existing mainstream knowledge. The LBD research (Petrič et al., 2012) and (Sluban et al., 2012) suggests that bridging terms are more frequent in documents that are in some sense different from the majority of documents in a given domain.

The outlier-based approach to LBD proposed by Petrič et al. (2012) uses document clustering to find outlier documents. The approach consists of two steps. In the first step, the OntoGen clustering algorithm (Fortuna, Grobelnik, and Mladenić, 2006) is applied to cluster the merged document set A ∪ C, consisting of documents from two domains A and C. The result of unsupervised clustering are two document clusters: A′ = *Classified as A* (i.e. documents from A ∪ C classified as A), and C′= *Classified as C* (i.e. documents from A ∪ C classified as C). In the second step of outlier detection, clusters A′ and C′ are further separated, each into two clusters, based on the documents' original labels A and C. As a result, a two-level tree hierarchy of clusters is generated, illustrated in Figure 2.

Lesson Learned 1: Potential of outlier documents. The hypothesis that outlier documents have the potential to improve the effectiveness of bridging term detection was tested on the *migraine-magnesium* (Swanson, Smalheiser, and Torvik, 2006) and *autism-calcineurin* (Petrič et al., 2009) domain pair datasets, which have lists of concept bridging terms (b-terms) confirmed by the medical experts. The experimental results obtained by using OntoGen confirm the hypothesis that most bridging terms appear in outlier documents and that by considering only outlier documents the search space for b-term identification can be largely reduced.

This lesson—that outlier documents have the potential for improving the effectiveness of bridging term detection—was reconfirmed in the work of Sluban et al. (2012), exploring a classification filtering approach to outlier detection, which was tested on the same domain pair data sets, *migraine-magnesium* (Swanson, Smalheiser, and Torvik, 2006) and *autism-calcineurin* (Petrič et al., 2009) domain, which have lists of bridging terms (b-terms) confirmed by the medical experts. Sluban et al. (2012) proposed to detect outlier documents using classification algorithms for classification noise filtering, first suggested by Brodley and Friedl (1999). Having documents from two domains of interest *A* and *C*, Sluban et al. (2012) first trained an ensemble classifier that distinguishes between the documents of these domains, and use the ensemble classifier to classify all the documents. The miss-classified documents were declared as outliers, since—according to the classification model—they do not belong to their domain (class label) of origin. These outliers can be interpreted as borderline documents as they were considered by the model to be more similar to the other domain

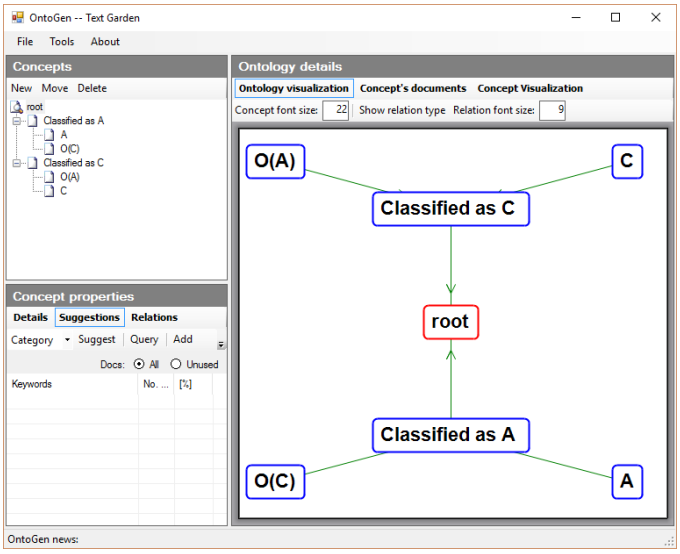

Figure 2: Target domain documents from disjoint literatures A and C, clustered according to the proposed OntoGen's two step approach, first using unsupervised and then supervised clustering to obtain outlier documents O(A) and O(C) of literatures A and C, respectively. The figure illustrates the outlier document detection approach as implemented in OntoGen, addressing the outlier detection framework, conceptually explained in Figure 3.

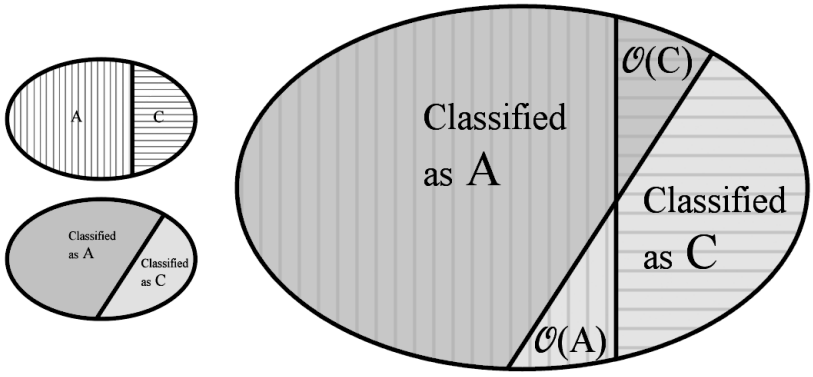

Figure 3: Detecting outliers of a domain pair dataset A ∪ C, using a document classification approach by Sluban et al. (2012).

than to their original domain, and can be regarded as bridging documents between the two domains. In other words, if an instance of class *A* is classified in the opposite class *C*, it is considered an outlier of domain *A*, and vice versa. The two sets of outlier documents were denoted with O(A) and O(C), illustrated in Figure 3.

The experimental results obtained by Sluban et al. (2012) showed that the sets of detected outlier documents are relatively small—including less than 5% of the entire datasets—and that they contain a great majority of bridging terms previously identified by medical experts, which was significantly higher than in same-sized random document subsets. These results are summarized in Figure 4.

These experimental results indicate that it is justified that the search for b-terms can be focused on outlier documents, which contain a large majority of b-terms. Consequently, by focusing the exploration on outlier documents, 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.

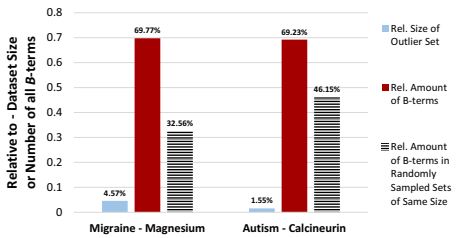

Figure 4: Presence of b-terms in the detected outlier sets of two domain pair datasets.

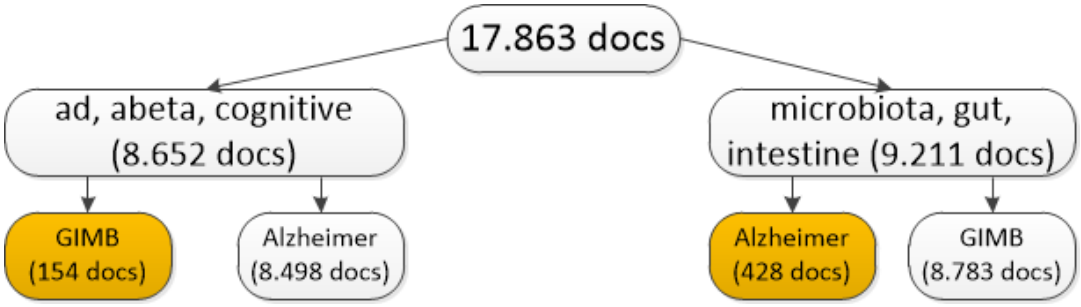

Figure 5: Two-level cluster hierarchy constructed with ontoGen from the dataset of 17,863 papers in the *Alzheimers disease-gut microbiome* domain pair.

Outlier-based LBD: Current applications and lessons learned

When applying OntoGen on the documents of the new application domain using the *Alzheimers disease-gut microbiome* domain pair (Cestnik et al., 2017), the OntoGen method uses domains A and C, and builds a joint document set $A \cup C$. With this intention, two individual sets of documents (e.g., titles, abstracts or full texts of scientific articles), one for each domain under research (namely, literature A on Alzheimer's disease and literature C on gut microbiome), were automatically retrieved from the PubMed database. A cluster hierarchy was constructed from the dataset of 17,863 papers with OntoGen. Two first-level clusters are labeled with the OntoGen suggested keywords *ad, abeta, cognitive* and *microbiota, gut, intestine*. Four second level sub-clusters separate documents according to their original search keywords for Alzheimers disease and gut microbiome, as illustrated in Figure 5.

Lesson Learned 2: Excluding intersecting documents. In the Alzheimers disease-gut microbiome LBD application, the initial document set $A \cup C$ consisted of some documents, which were in the intersection of A and C, meaning that a few documents were retrieved from PubMed by both of the two separate queries for domain A (i.e. *Alzheimer*) and C (i.e. (*gut* OR *intestinal*) AND (*microbiota* OR *bacteria*)), which was surprising. After carefully inspecting these documents (as these documents could contain the b-terms representing a solution to the problem, which proved not to be the case) it was realized that keeping them in the $A \cup C$ document set was problematic. As a result, the documents that were retrieved by both queries were eliminated[2], resulting in 17,863 documents kept in the $A \cup C$ document set used for further exploration.

Lesson Learned 3: Selecting only outlier documents. The hypothesis that the search for bridging terms can be reduced to manageable subsets of documents was confirmed in our experiments. In the Alzheimers disease-gut microbiome LBD application using OntoGen for outlier document detection, the space of documents used for b-term exploration was further reduced from the set of 17,863 documents to two subsets of outlier documents, i.e. to only 154 gut microbiome papers and 428 Alzheimer's disease related papers, considered as outliers in their own domain, leading to the selection of only 582 documents for further inspection.

Lesson Learned 4: Expert revision of b-terms list. The hypothesis that b-terms selected from outlier documents can be further reduced with expert knowledge was confirmed in our experiments. By processing the remaining 582 outlier documents, we used CrossBee (Juršič et al., 2012) to extract 4,723 terms as potential b-terms connecting the two domains. In b-term exploration all the terms were considered and not just the medical ones, except that a list of 523 English stop words was used to filter out meaningless words, and English Porter stemming was applied, which helped us to focus on medically interesting b-terms. Even though the list of potential bridging terms was ordered according to the ensemble-heuristics estimated bridging terms potential, browsing and analyzing the terms from the list still presented a substantial burden for the domain expert. To further reduce the size of the potential b-term list, the collaborating domain expert[3] prepared a list of 289 domain terms of her own research interest. This list included common terms and specific molecular factors and pathways, which were manually identified in titles, abstracts, and keywords from 42 papers obtained from PubMed search query (*gut AND Alzheimer*), 55 of which appeared also among the 4,723 terms extracted by CrossBee. During the evaluation phase, the relevant papers for each b-term candidate were reviewed and searched for potential clues justifying further investigation, resulting from relevant b-term discoveries confirmed by the domain expert (Cestnik et al., 2017).

Compared to outlier document detection using OntoGen, an upgraded methodology proposed by Cestnik et al. (2017) was implemented in a reusable outlier based LBD methodology in a web-based text mining platform TextFlows[4]

[2]Their inclusion in the document set would have violated the assumption of literature-based discovery and bisociative knowledge discovery frameworks, which assume that the explored literature domains A and C are disjoint; if this assumption were violated, the methodology would fail due to biased heuristics calculations.

[3]Elsa Fabretti

[4]http://textflows.org

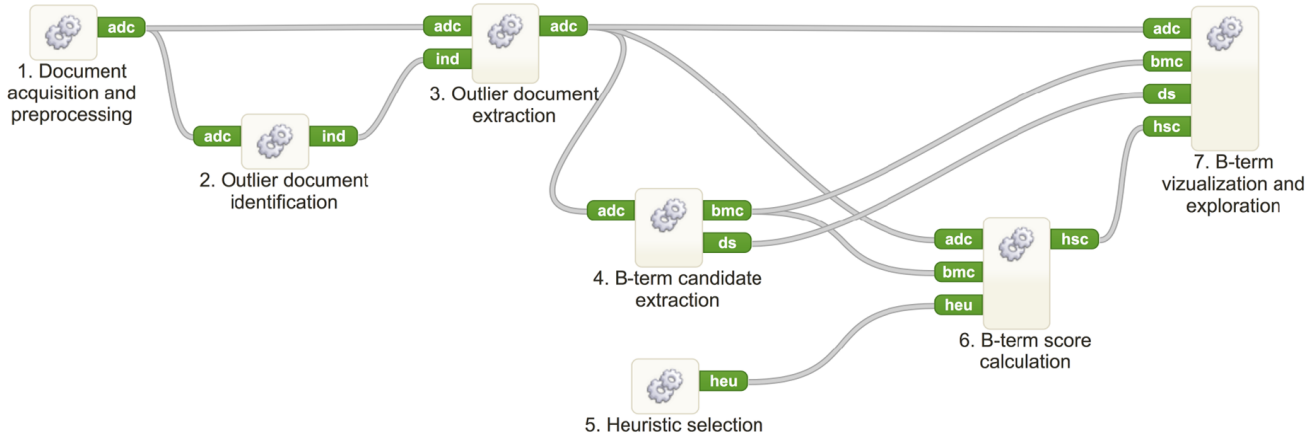

Figure 6: A top-level workflow of the LBD methodology in TextFlows (Perovšek et al., 2016).

(Perovšek et al., 2016) that allowed us to construct and execute advanced text mining workflows. The workflow shown in Figure 6 consists of seven steps implemented as sub-processes. The connections between sub-processes represent the flow of documents from one sub-process to another. In overview, steps 1-3 represent the outlier detection part, and steps 4-7 represent cross-domain exploration for b-term detection.

Lesson Learned 5: TextFlows workflow helping experts. In the experiments using the TextFlows workflow, the NoiseRank ensemble-based outlier detection approach (Sluban, Gamberger, and Lavrač, 2013) as implemented in TextFlows was used. The goal of the first three steps (using first three workflow widgets) of the methodology is to effectively extract a set of outlier documents from the whole corpus of input documents. Consequently, by decreasing the size of the input set of documents the second phase becomes more focused, efficient and effective. In the last four steps of the workflow in Figure 6 components that constitute the CrossBee HCI interface (Juršič et al., 2012) are executed to conduct expert-guided b-term analysis. Here, the goal is to further prepare the input documents for b-term visualization and exploration. Note that in this step the role of the domain expert is crucial.

Lesson Learned 6: Term filtering and synonyms matter. In recent LBD experiments, using *plant defence-circadian rhythm* domain pair, the goal was to identify potentially interesting new daily regulated mechanisms that are responsible for plant defence. After obtaining 5,412 documents from PubMed containing complete articles (2,483 from plant defence and 2,929 from circadian rhythm), 0.5% documents shorter than 20 characters (mostly empty contents) and longer than 97,500 characters (containing many different articles in proceedings) were removed. Then, 12 duplicates that were present in both domains (as in Lesson Learned 2) were eliminated. The crucial, although simple and straightforward, step in this experiment was the replacement of gene names with synonyms gathered in previous research projects (22,265 gene names mapped into 7,863 synonyms). In addition, the documents were optionally pre-processed to keep only gene-related terms (included in synonym list and from the gene dictionary containing additional 6,083 gene names), which resulted in a substantial reduction of the input file size (from 200 MB to 28 MB).

Current research lessons

Future work, aimed at improving the effectiveness of bridging term detection in cross-domain literature mining, will be performed in several directions. It will be based on the lessons learned in the current research: using embeddings for representation learning used in document clustering, using ontologies for term enrichment in cross-domain document exploration, and using network analysis for cross-domain heterogeneous information network exploration.

Related Lesson 1. The use of background knowledge remains largely unexploited in text classification and clustering. Word taxonomies can easily be exploited as means for constructing new semantic features, which can be used in text representation learning to improve the performance and robustness of the learned models. Consequently, recently developed tax2vec algorithm could be used for constructing taxonomy based features to improve the results of document clustering and classification.

Related Lesson 2. Given that documents can be easily transformed into graphs (e.g., graphs constructed from subject-verb-object triplets), network analysis approaches can prove to be fruitful for bridging term detection (e.g., community detection and finding bridging nodes in graphs between subgraphs representing the detected communities). In addition to network analysis approaches, novel graph embedding approaches could also be used in this context.

Related Lesson 3. Instead of using the standard TF-IDF (Term Frequency Inverse Document Frequency) weighted Bag Of Words vector representations of text documents, which was used in the past LBD research outlined in this paper (Petrič et al., 2012; Sluban et al., 2012; Juršič et al., 2012; Cestnik et al., 2017), the current research in EMBEDDIA[5] indicates that representation learning using embeddings is much more effective than using the standard TF-IDF Bag of Words document representation. Consequently, improved clustering results can be expected using contemporary embedding approaches such as word2vec, doc2vec or Bert.

Complying with Related Lesson 3, this section proposes a novel approach to bisociative discovery between two separate domains *A* and *C*, using the power of word embeddings.

Word embeddings are vector representations of words: each word is assigned a vector of several hundred dimensions. These are usually obtained via training algorithms such as word2vec (Mikolov et al., 2013), GloVe (Pennington, Socher, and Manning, 2014) or FastText (Bojanowski et al., 2017), which characterize the word based on the lexical context in which it appears. These representations improve performance in a wide range of automated text processing tasks, partly because they capture a degree of semantics. They can also capture regularities beyond simple relatedness, such as analogies (Mikolov, Yih, and Zweig, 2013); for example, the vector-space relation between Madrid and Spain is very similar to that between Paris and France.

In a closed literature based discovery setting, we are interested if a specific relation between two concepts ($a1$ and $a2$) in the first domain *A* could also be found between concepts *x* and *c* in the second domain *C*, where concept *c* is given in advance and *x* is the new concept that we are trying to find.

More formally, this can be written in a form of an analogy (i.e. bisociation) between two separate domains *A* and *C*:

```
a1 rel a2 == x rel c
```

In the embedding space, this analogy translates to the following equation between embeddings:

```
x = embedding a1 + embedding a2 -
            embedding c
```

Finally, once *x* is calculated, we need to find a set of concepts from the second domain that have an embedding representation most similar to *x* according to the cosine similarity.

In the context of computational creativity research based on bisociation (Koestler, 1964), bisociative patterns that are searched and explored include: bridging concepts, bridging graphs, and bridging by structural similarity (Kötter and Berthold, 2012). The embeddings-based bisociative knowledge discovery approach described above addresses the latter, most complex setting of bridging by structural similarity, defined as follows:

Bridging by structural similarity. This is the most complex kind of bisociation, whereby in a bisociative network representation of concepts, subsets of concepts in each domain share structural similarities. Bisociations based on structural similarity are represented by relations and/or sub-graphs of two different, structurally-similar domains Kötter and Berthold (2012).

This type of bisociation is according to Kötter and Berthold (2012) the most abstract pattern with the potential for new cross-domain discoveries, which e.g., graph similarity methods can identify.

In our preliminary experiments using *plant defence-circadian rhythm* domain pair, where the goal was to identify potentially interesting new daily regulated mechanisms that are responsible for plant defence, we employed FastText embeddings (Bojanowski et al., 2017), in which a word is represented as an average of its character n-grams. This allows the model to leverage both semantic and morphological information, which is useful in a setting with small domain corpora containing less semantic information, since morphological similarity in many cases translates to semantic relatedness. Separate embedding models were trained for domains *A* and *C* and then aligned into a common vector space by using a supervised approach that relies on a training dictionary of identical words from both domains, used as anchor points to learn a mapping from the source to the target space with Procrustes alignment (Conneau et al., 2017).

Conclusions

This paper addresses the field of scientific computational creativity, in particular bisociative literature-based discovery. The paper mostly focusing on finding outlier documents as means for finding unexpected links crossing different contexts. Selected approaches to bridging term detection through outlier document exploration are briefly outlined, together with the lessons learned from recent applications in medical and biological literature-based knowledge discovery. Finally, the paper addresses new prospects in bisociative literature-based discovery, emphasizing the use of advanced embeddings technology for cross-domain literature mining.

In future work we will further explore embeddings-based LBD both in the closed and in the open LBD settings. We will also introduce additional user interface options for data visualization and exploration as well as advance the term ranking methodology by adding new sophisticated heuristics, which will take into account also the semantic aspects of the data. Finally, by using the recent word embedding technology, we aim to implement a novel bisociative knowledge discovery setting of bridging by structural similarity.

Acknowledgements

This work was supported by Slovenian Research Agency grant P-0103 and the European Unions Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. I.; et al. 1996. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining* 12(1):307–328.

[5]See acknowledgements.

Berthold, M., ed. 2012. *Bisociative Knowledge Discovery*. Springer.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135146.

Brodley, C. E., and Friedl, M. A. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research* 11:131–167.

Bruza, P., and Weeber, M. 2008. *Literature-based Discovery*. Springer Science & Business Media.

Cestnik, B.; Fabbretti, E.; Gubiani, D.; Urbančič, T.; and Lavrač, N. 2017. Reducing the search space in literature-based discovery by exploring outlier documents: A case study in finding links between gut microbiome and alzheimer's disease. *Genomics and Computational Biology* 3(3):e58.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word translation without parallel data. *CoRR* abs/1710.04087.

Fortuna, B.; Grobelnik, M.; and Mladenić, D. 2006. Semi-automatic data-driven ontology construction system. In *Proceedings of the 9th International Multi-conference Information Society*, 223–226.

Gopalakrishnan, V.; Jha, K.; Jin, W.; and Zhang, A. 2019. A survey on literature based discovery approaches in biomedical domain. *Journal of Biomedical Informatics* 93:103141.

Holzinger, A.; Yildirim, P.; Geier, M.; and Simonic, K.-M. 2013. Quality-based knowledge discovery from medical text on the web. In *Quality Issues in the Management of Web Information*. Springer. 11–13.

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

Juršič, M.; Cestnik, B.; Urbančič, T.; and Lavrač, N. 2012. Cross-domain literature mining: Finding bridging concepts with CrossBee. In *Proceedings of the 3rd International Conference on Computational Creativity*, 33–40.

Kastrin, A.; Rindflesch, T. C.; and Hristovski, D. 2014. Link prediction on the semantic MEDLINE network. In *Proceedings of the International Conference on Discovery Science*, 135–143. Springer.

Koestler, A. 1964. *The Act of Creation*. Hutchinson.

Kötter, T., and Berthold, M. 2012. From information networks to bisociative information networks. In *Bisociative Knowledge Discovery*, 33–50. Springer.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *Proceedings of ICLR* CoRR. abs/1301.3781.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics.

Perovšek, M.; Kranjc, J.; Erjavec, T.; Cestnik, B.; and Lavrač, N. 2016. TextFlows: A visual programming platform for text mining and natural language processing. *Science of Computer Programming* 121:128–152.

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

Petrič, I.; Cestnik, B.; Lavrač, N.; and Urbančič, T. 2012. Outlier detection in cross-context link discovery for creative literature mining. *The Computer Journal* 55(1):47–61.

Sebastian, Y.; Siew, E.-G.; and Orimaye, S. O. 2017. Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review* 32:e12.

Sluban, B.; Juršič, M.; Cestnik, B.; and Lavrač, N. 2012. Exploring the power of outliers for cross-domain literature mining. In *Bisociative Knowledge Discovery*, 325–337. Springer.

Sluban, B.; Gamberger, D.; and Lavrač, N. 2013. Ensemble-based noise detection: Noise ranking and visual performance evaluation. *Data Mining and Knowledge Discovery* 1–39.

Smalheiser, N., and Swanson, D. R. 1998. Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine* 57(3):149–154.

Swanson, D. R.; Smalheiser, N. R.; and Torvik, V. I. 2006. 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.

Swanson, D. R. 1990. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association* 78(1):29.

Weeber, M.; Klein, H.; de Jong-van den Berg, L.; Vos, R.; et al. 2001. 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):548–557.

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